Flexible Covariate Adjustments in Regression Discontinuity Designs

Claudia Noack Tomasz Olma Christoph Rothe

Abstract

Empirical regression discontinuity (RD) studies often use covariates to increase the precision of their estimates. In this paper, we propose a novel class of estimators that use such covariate information more efficiently than the linear adjustment estimators that are currently used widely in practice. Our approach can accommodate a possibly large number of either discrete or continuous covariates. It involves running a standard RD analysis with an appropriately modified outcome variable, which takes the form of the difference between the original outcome and a function of the covariates. We characterize the function that leads to the estimator with the smallest asymptotic variance, and show how it can be estimated via modern machine learning, nonparametric regression, or classical parametric methods. The resulting estimator is easy to implement, as tuning parameters can be chosen as in a conventional RD analysis. An extensive simulation study illustrates the performance of our approach.

First Version: July 16, 2021. The authors gratefully acknowledge financial support by the European Research Council (ERC) through grant SH1-77202. Author contact information: Claudia Noack, Department of Economics, University of Mannheim. E-Mail: cnoack@mail.uni-mannheim.de. Website: https://claudianoack.github.io. Tomasz Olma, Department of Economics, University of Mannheim. E-Mail: tolma@mail.uni-mannheim.de. Website: https://tomaszolma.github.io. Christoph Rothe, Department of Economics, University of Mannheim. E-Mail: rothe@vwl.uni-mannheim.de. Website: http://www.christophrothe.net.

1. INTRODUCTION

Regression discontinuity (RD) designs are widely used for estimating causal effects from observational data in economics and other social sciences. These designs exploit that in many contexts a unit's treatment status is determined by whether its realization of a running variable exceeds some known cutoff value. For example, students might qualify for a scholarship if their GPA is above some threshold. Under continuity conditions on the distribution of potential outcomes, the average treatment effect at the cutoff is identified in such designs by the size of the jump in the conditional expectation of the outcome given the running variable at the cutoff. Estimation and inference is then commonly based on local linear regression techniques (e.g., Hahn et al., 2001; Imbens and Kalyanaraman, 2012; Calonico et al., 2014; Armstrong and Kolesár, 2020).

An RD analysis generally does not require data beyond the outcome and the running variable, but in practice researchers often use additional covariates, such as socio-demographic characteristics, to reduce the variance of their empirical estimates. A common strategy is to include the covariates additively separably and linearly-in-parameters in a local linear RD regression (Calonico et al., 2019). Such linear adjustment estimators are consistent without functional form assumptions on the underlying conditional expectations if the covariates are predetermined, but generally do not exploit the available covariate information efficiently. They are also not able to handle settings in which the number of covariates is large, which is common when working, for example, with administrative data.

In this paper, we propose a novel class of covariate-adjusted RD estimators that combine linear regression methods with flexible covariate adjustments, possibly based modern machine learning techniques, to address these issues. To motivate the approach, let Y_i and Z_i denote the outcome and covariates, respectively, of observational unit i. Calonico et al. (2019) show that linear adjustment estimators are asymptotically equivalent to local linear RD regressions with the modified outcome variable $Y_i - Z_i^\top \gamma$, where γ is a vector of projection coefficients. We consider generalizations of such estimators with a modified outcome of the form $Y_i - \mu(Z_i)$, for some generic function μ.

Such estimators are easily seen to be consistent for *any* fixed μ if the distribution of the covariates varies smoothly around the cutoff in some appropriate sense, which is compatible with the notion of covariates being "predetermined". We also show that their asymptotic variance is minimized if $\mu = \mu_0$ is the average of the two conditional expectations of the outcome variable given the running variable and the covariates just above and below the cutoff. This optimal adjustment function is generally nonlinear and not known in practice, but can be estimated from the data.

Our proposed estimators hence take the form of a local linear RD regression with the generated outcome $Y_i - \widehat{\mu}(Z_i)$, where $\widehat{\mu}$ is some estimate of μ_0 obtained in a preliminary stage. We implement such estimators with cross-fitting (e.g., Chernozhukov et al., 2018), which is an efficient form of sample splitting that removes certain biases and allows us to accommodate a wide range of estimators of the optimal adjustment function. In particular, one can use modern machine learning methods like lasso regression, random forests, deep neural networks, or ensemble combinations thereof, to estimate the optimal adjustment function. However, researchers can also use classical nonparametric approaches like local polynomials or series regression; or estimators based on fully parametric specifications.

Importantly, valid inference on the RD parameter in our setup does not require that μ_0 is consistently estimated: our theory only requires that the first-stage estimates concentrate in a mean-square sense around some deterministic function $\bar{\mu}$, which could in principle be different from μ_0, but the rate of this convergence could be arbitrarily slow. Our setup can allow for this because our proposed RD estimators are very insensitive to estimation errors in the preliminary stage. This is because they are constructed as sample analogues of a moment function that contains μ_0 as a nuisance function, but does not vary with it: as discussed above, our parameter of interest is equal to the jump in the conditional expectation of $Y_i - \mu(Z_i)$ given the running variable at the cutoff for *any* fixed function μ. This insensitivity property is related to Neyman orthogonality, which features prominently in many modern two-stage estimation methods (e.g., Chernozhukov et al., 2018), but it is a global rather than a local property and is thus in effect substantially stronger.[1]

Our theoretical analysis specifically shows that, under the conditions outlined above, our proposed RD estimator is first-order asymptotically equivalent to a local linear RD estimator with $Y_i - \bar{\mu}(Z_i)$ as the dependent variable. This result is then used to study its asymptotic bias and variance, and to derive an asymptotic normality result. The asymptotic variance of our estimator depends on the function $\bar{\mu}$ and achieves its minimum value if $\bar{\mu} = \mu_0$ (that is, if μ_0 is consistently estimated in the first stage). We also propose a standard error that is valid in either case. Note that since our result does not require a particular rate of

[1] A moment function is Neyman orthogonal if its first functional derivative with respect to the nuisance function is zero, but the (conditional) moment function on which our estimates are based is fully invariant with respect to the nuisance function. Chernozhukov et al. (2018) give several examples of setups in which such a property occurs, which include optimal instrument problems, certain partial linear models, and treatment effect estimation under unconfoundedness when the propensity score is known. Such global insensitivity is also easily seen to occur more generally in setups with doubly robust (cf. Robins and Rotnitzky, 2001) moments if one of the nuisance function is known.

convergence for the first step estimate of μ_0, our RD estimator is largely unaffected by "curse of dimensionality", and can be expected to perform well in settings with many covariates.

Practical issues like bandwidth choice and construction of confidence intervals with good coverage properties are also rather straightforward in our setting. In particular, our results justify simply apply existing methods to a data set in which the outcome Y_i is replaced with the generated outcome $Y_i - \widehat{\mu}(Z_i)$, ignoring that $\widehat{\mu}$ has been estimated. This can easily be accomplished using existing software packages. In an extensive simulation study, we also show that our theoretical findings provide very good approximations to our estimators' finite sample behavior.

These types of results are qualitatively similar those that have been obtained for efficient influence function (EIF) estimators of the population average treatment effect in simple randomized experiments with known and constant propensity scores (e.g., Wager et al., 2016). Parallels arise because EIF estimators are also based on a moment function that is globally invariant with respect to a nuisance function. In fact, we argue that our RD estimator is in many ways a direct analogue of the EIF estimator, and the variance that it achieves under the optimal adjustment function is also similar in structure to the semiparametric efficiency bound in simple randomized experiments.

Related Literature. Our paper contributes to an extensive literature on estimation and inference in RD designs; see, e.g., Imbens and Lemieux (2008) and Lee and Lemieux (2010) for a surveys, and Cattaneo et al. (2019) for a textbook treatment. Different ad-hoc methods for incorporating covariates into an RD analysis have long been used in applied economics (see, e.g., Lee and Lemieux, 2010, Section 3.2.3). Following Calonico et al. (2019), it has become common practice to include covariates without localization into the usual local linear regression estimator. We show that our approach nests this estimator as a special case, but is generally more efficient. Other closely related papers are Kreiß and Rothe (2021), who extend the approach in Calonico et al. (2019) to settings with high-dimensional covariates under sparsity conditions, and Frölich and Huber (2019), who propose to incorporate covariates into an RD analysis in a fully nonparametric fashion. The latter method is generally quite affected from the curse of dimensionality, and is thus unlikely to perform well in practice.

Our paper is also related in a more general way to the vast literature on two-step estimation problems with infinite-dimensional nuisance parameters (e.g., Andrews, 1994; Newey, 1994), especially the recent strand that exploits Neyman orthogonal (or debiased) moment functions and cross-fitting (e.g., Belloni et al., 2017; Chernozhukov et al., 2018). The latter literature focuses mostly on regular (root-n estimable) parameters, while our RD treatment effect is a non-regular (nonparametric) quantity. Some general results on non-regular estimation based on orthogonal moments are derived in Chernozhukov et al. (2019), and specific results for estimating conditional average treatment effects in models with unconfoundedness are given, for example, in Kennedy et al. (2017), Kennedy (2020) and Fan et al. (2020). Our results differ from those in these papers because, as explained above, our estimator is based on a moment function that satisfies a property that is stronger than orthogonality.

Plan of the Paper. The remainder of this paper is organized as follows. In Section 2, we introduce the setup. In Section 3, we describe our proposed covariate-adjusted RD estimator. In Section 4, we present our main theoretical results. Further results are discussed in Section 5. Section 6 contains a simulation study. Section 7 concludes.

2. SETUP

In this section, we introduce the model and the parameter of interest. Furthermore, we discuss estimation of the RD parameter based on local linear regression methods.

2.1. **Model and Parameter of Interest.**

We begin by considering sharp RD designs; see Section 5.3 for a discussion of fuzzy RD setups. Our object of interest is the causal effect of a binary treatment on some outcome variable. The data $\{W_i\}_{i\in[n]} = \{(Y_i, X_i, Z_i)\}_{i\in[n]}$, where $[n] = \{1, \ldots, n\}$, an i.i.d. sample of size n from the distribution of $W = (Y, X, Z)$. Here, $Y_i \in \mathbb{R}$ is the outcome variable, $X_i \in \mathbb{R}$ is the running variable, and $Z_i \in \mathbb{R}^d$ is a vector of covariates. Units receive the treatment if and only if the running variable exceeds some known threshold, which we normalize to zero without loss of generality. We denote the treatment indicator by T_i, so that $T_i = \mathbf{1}\{X_i \geq 0\}$. The parameter of interest is the height of the jump in the conditional expectation of the observed outcome variable given the running variable at zero:

$$\tau = \mathbb{E}[Y_i|X_i = 0^+] - \mathbb{E}[Y_i|X_i = 0^-], \tag{2.1}$$

where we use the notation that $f(0^+) = \lim_{x\downarrow 0} f(x)$ and $f(0^-) = \lim_{x\uparrow 0} f(x)$ are the right and left limit, respectively, of a generic function $f(x)$ at zero. In a potential outcomes framework, this parameter coincides with the average treatment effect of units at the cutoff under standard continuity conditions (Hahn et al., 2001), but we consider estimation of τ as defined in (2.1).

Throughout the paper, we assume that the distribution of the running variable X_i is fixed, but we allow the conditional distribution of (Y_i, Z_i) given X_i to change with the sample size in our asymptotic analysis. In particular, we allow the dimension of Z_i to grow with n, but we generally leave such dependence on n implicit in our notation.

2.2. **Standard RD Estimator.** In RD designs without covariates the parameter of interest is typically estimated by running a local linear regression (Fan and Gijbels, 1996) on each side of the cutoff. That is, the estimator takes the form

$$\widehat{\tau}(h) = e_1^\top \operatorname*{argmin}_{\beta \in \mathbb{R}^4} \sum_{i=1}^n K_h(X_i)(Y_i - (T_i, X_i, T_i X_i, 1)^\top \beta)^2, \tag{2.2}$$

where $K(\cdot)$ is a kernel function with support $[-1, 1]$, $h > 0$ is a bandwidth, $K_h(v) = K(v/h)/h$, and $e_1 = (1, 0, 0, 0)^\top$ is the first unit vector. This estimator can also be written as a weighted sum of the realizations of the outcome variable,

$$\widehat{\tau}(h) = \sum_{i=1}^n w_i(h) Y_i,$$

with weights $w_i(h)$ that depend on the data through the realizations of the running variable only; see Appendix A.1 for an explicit expression. Under standard assumptions, which include that the running variable is continuously distributed, and that the bandwidth tends to zero at an appropriate rate, the estimator $\widehat{\tau}(h)$ is approximately normally distributed in large samples, with bias of order h^2 and variance of order $(nh)^{-1}$:

$$\begin{aligned}\widehat{\tau}(h) \overset{a}{\sim} N \Big(&\tau + h^2 \frac{\bar{\nu}}{2} \left(\partial_x^2 \mathbb{E}[Y_i | X_i = x]|_{x=0^+} - \partial_x^2 \mathbb{E}[Y_i | X_i = x]|_{x=0^-} \right), \\ &\frac{1}{nh} \frac{\bar{\kappa}}{f_X(0)} (\mathbb{V}[Y_i | X_i = 0^+] + \mathbb{V}[Y_i | X_i = 0^-]) \Big),\end{aligned} \tag{2.3}$$

where $\bar{\nu}$ and $\bar{\kappa}$ are kernel constants, and f_X denotes the density of X_i.[2]

3. COVARIATE-ADJUSTED RD ESTIMATOR

In this section, we describe our proposed procedure. We begin with a brief overview and then motivate and describe the estimator more formally. We also discuss its relationship to existing procedures, and its analogy to efficient estimation in randomized experiments.

[2] The continuity of the running variable's density f_X around the cutoff that is implied by this expression is strictly speaking not needed to derive a normality result like (2.3). However, a jump in f_X at the cutoff is typically taken as an indication that the assumptions that justify interpreting τ as a causal parameter are unlikely to be satisfied (McCrary, 2008; Gerard et al., 2020). We therefore focus on the case where f_X is continuous in this paper.

3.1. **Overview.** We first give a brief overview of our proposed procedure and the theoretical results we derive for it. The general idea is to run a conventional sharp RD estimator with a modified dependent variable that is estimated from the data, using cross-fitting to allow for a wide range of methods. Specifically, this involves the following two steps:

1. Randomly split the data $\{(Y_i, X_i, Z_i)\}_{i\in[n]}$ into S folds of (roughly) equal size. For $s \in [S]$, let $\widehat{\mu}_s^+(z)$ and $\widehat{\mu}_s^-(z)$ be generic estimates of $\mathbb{E}[Y_i|X_i = 0^+, Z_i = z]$ and $\mathbb{E}[Y_i|X_i = 0^-, Z_i = z]$ that use all the data but that in the sth fold. These estimates could be obtained via simple parametric approaches like linear regression, classical semi-nonparametric regression techniques, or modern machine learning methods.

2. Put $\widehat{\mu}_{s(i)}(Z_i) = (\widehat{\mu}_{s(i)}^+(Z_i) + \widehat{\mu}_{s(i)}^-(Z_i))/2$, where $s(i)$ denotes the fold that contains observation i, write $M_i(\mu) = Y_i - \mu(Z_i)$, and estimate τ by

$$\widehat{\tau}_{CF}(h; \widehat{\mu}) = \sum_{i=1}^{n} w_i(h) M_i(\widehat{\mu}_{s(i)}). \tag{3.1}$$

We motivate this estimator in Section 3.2 as a generalization of the linear adjustments in Calonico et al. (2019), and show in Section 3.6 that the latter are indeed a special case of our approach (except for cross-fitting). In Section 5.2, we also show that our estimator is conceptually analogous to an efficient influence function estimator of the population average treatment effect in a randomized experiment with constant propensity score.

Our theoretical analysis in Section 4 then establishes that, under rather weak regularity conditions, the estimator $\widehat{\tau}_{CF}(h; \widehat{\mu})$ is asymptotically equivalent to a version of it that uses the variable $M_i(\bar{\mu})$ as the outcome, where $\bar{\mu}(Z_i) = (\bar{\mu}^+(Z_i) + \bar{\mu}^-(Z_i))/2$, and $\bar{\mu}^+(z)$ and $\bar{\mu}^-(z)$ are deterministic approximations of $\widehat{\mu}_s^+(z)$ and $\widehat{\mu}_s^-(z)$, respectively, whose error vanishes in large samples in some appropriate sense. In view of (2.3), it thus holds that

$$\begin{aligned}\widehat{\tau}_{CF}(h; \widehat{\mu}) \overset{a}{\sim} N \Big(&\tau + h^2 \frac{\bar{\nu}}{2} \left(\partial_x^2 \mathbb{E}[M_i(\bar{\mu})|X_i = x]|_{x=0^+} - \partial_x^2 \mathbb{E}[M_i(\bar{\mu})|X_i = x]|_{x=0^-}\right), \\ &\frac{1}{nh} \frac{\bar{\kappa}}{f_X(0)} (\mathbb{V}[M_i(\bar{\mu})|X_i = 0^+] + \mathbb{V}[M_i(\bar{\mu})|X_i = 0^-])\Big).\end{aligned}$$

We also show that the asymptotic variance in this last equation is minimized if $\bar{\mu}^+(z)$ and $\bar{\mu}^-(z)$ coincide with $\mathbb{E}[Y_i|X_i = 0^+, Z_i = z]$ and $\mathbb{E}[Y_i|X_i = 0^-, Z_i = z]$, respectively, but the above approximation still holds if that is not the case. Bandwidth choice and inference can thus be carried out by applying standard methods to an RD design with outcome $M_i(\widehat{\mu}_{s(i)})$ and ignoring the sampling uncertainty about the estimate of μ_0.

3.2. **Motivating Covariate-Adjusted Outcome Variables.** It is common practice in empirical work to include additional pretreatment covariates linearly and without localization into the regression (2.2). While such linear adjustments typically reduce the variance of the final RD estimator, there are potentially more efficient ways to exploit covariate information for inference. To motivate our procedure, note that Calonico et al. (2019) show a linear adjustment estimator is asymptotically equivalent to an RD estimator of the form in (2.2) that uses the modified outcome variable $Y_i - \gamma^\top Z_i$ instead of Y_i, with γ a certain vector of linear projection coefficients. Our approach is to directly implement an RD estimator with a covariate-adjusted outcome, but for a much more general class of adjustments. That is, one could consider estimators of the form

$$\widehat{\tau}(h;\mu) = \sum_{i=1}^{n} w_i(h) M_i(\mu), \quad M_i(\mu) = Y_i - \mu(Z_i), \tag{3.2}$$

where μ is a function to which we refer in the following as the adjustment function.

One of our key assumptions ensures that such estimators are consistent. Intuitively, it states that the conditional distribution of the covariates given the running variable changes smoothly around the cutoff; see Section 4 for a precise statement. This condition formalizes the idea of the covariates being determined prior to treatment assignment, which is also needed with other approaches to covariate adjustment (Calonico et al., 2019). It implies that $\mathbb{E}\left[\mu(Z_i)|X_i = x\right]$ is continuous in x on $\mathcal{X}$ for all n and all suitably integrable functions μ, which in turn implies that

$$\tau = \mathbb{E}[M_i(\mu)|X_i = 0^+] - \mathbb{E}[M_i(\mu)|X_i = 0^-] \tag{3.3}$$

for all such suitable functions μ. That is, we can exchange the outcome variable Y_i in the definition of τ in (2.1) for $M_i(\mu)$ without affecting the validity of the equation. The term on the right-hand-side of (3.3) is a moment function that identifies τ, but is globally insensitive to variation in μ. This property is key for our theoretical results below.

3.3. **Optimal Adjustment Function.** In view of (2.3), the leading bias of the estimator $\widehat{\tau}(h;\mu)$ depends on μ only through a term that is proportional to

$$\partial_x^2 \mathbb{E}[\mu(Z_i)|X_i = x]|_{x=0^+} - \partial_x^2 \mathbb{E}[\mu(Z_i)|X_i = x]|_{x=0^-}.$$

With our assumption that the distribution of the covariates varies smoothly around the cutoff, we expect this term to show little variation in μ, or even not to depend on μ at all.

We therefore consider choosing the function μ in (3.2) such that the asymptotic variance of $\widehat{\tau}(h;\mu)$ is as small as possible. Again in view of (2.3), to achieve this, it suffices to minimize

$$\mathbb{V}[M_i(\mu)|X_i = 0^+] + \mathbb{V}[M_i(\mu)|X_i = 0^-]. \tag{3.4}$$

Our assumption that the conditional distribution of the covariates given the running variable changes smoothly around the cutoff also implies that $\mathbb{V}\left[\mu(Z_i)|X_i = x\right]$ is continuous in x on $\mathcal{X}$ for all n and μ. Simple algebra then shows that (3.4) is minimized by the function

$$\mu_0(z) = \frac{1}{2}\left(\mu_0^+(z) + \mu_0^-(z)\right), \quad \mu_0^\star(z) = \mathbb{E}[Y_i|X_i = 0^\star, Z_i = z] \text{ for } \star \in \{+,-\}. \tag{3.5}$$

That is, the optimal adjustment function μ_0 is the equally-weighted average of the left and right limit of the “long” conditional expectation function $\mathbb{E}[Y_i|X_i = x, Z_i = z]$ at the cutoff value $x = 0$. To see this, let $R(\mu) = \mathbb{V}[\mu_0^+(Z_i) - \mu(Z_i)|X_i = 0] + \mathbb{V}[\mu_0^-(Z_i) - \mu(Z_i)|X_i = 0]$, and note that (3.4) is equal to

$$\mathbb{V}[M_i(\mu_0^+)|X_i = 0^+] + \mathbb{V}[M_i(\mu_0^-)|X_i = 0^-] + R(\mu).$$

Since the first two terms on the right-hand side of the previous equation do not depend on μ, it suffices to minimize $R(\mu)$. The first and second summand in the definition of $R(\mu)$ could be set to zero separately by choosing μ as μ_0^+ or μ_0^-, respectively. Since

$$R(\mu) = R(\mu_0) + 2\mathbb{V}[\mu_0(Z_i) - \mu(Z_i)|X_i = 0] \geq R(\mu_0),$$

it turns out that the full sum is minimized by the average μ_0 of these two functions. Note that (3.4) is also minimized by $c + \mu_0$ for any constant $c \in \mathbb{R}$.[3]

3.4. **Estimator.** We propose to estimate the RD parameter τ by a feasible version of the estimator $\widehat{\tau}(h;\mu_0)$ based on a preliminary estimate of the function μ_0 defined in (3.5). Typically this preliminary estimate will be obtained by averaging separate estimates of μ_0^+ and μ_0^-, which can be obtained by any method deemed suitable for estimating such conditional expectations in the respective empirical context. Given the large data sets commonly encountered in empirical RD applications these days, modern machine learning tools are an attractive option for this task. However, researchers could also opt for simple parametric approaches like linear regression, or classical nonparametric regression (we discuss this in

[3] We obtain the same result if we consider the asymptotic variance of the estimator $\widehat{\tau}(h(\mu);\mu)$, with $h(\mu)$ the bandwidth that minimizes a performance criterion that involves a first-order bias-variance trade-off, like the asymptotic mean squared error.

more detail in the next subsection).

Cross-fitting (see, e.g., Chernozhukov et al., 2018) allows us to accommodate any of the just-mentioned methods. It is an efficient form of sample splitting that removes certain biases, and allows for a theoretical analysis that requires only weak assumptions about the preliminary estimate of μ_0. Cross-fitting would not be required for simple first-stage estimators, such as those based on linear regression specifications of μ_0^+ or μ_0^-, but we still maintain it to keep a unified treatment.

To explain the cross-fitting procedure, let $\widehat{\mu}(z; \{W_i\}_{i\in[n]})$ be the researcher's preferred estimator of μ_0, calculated on the full sample. We then split the data randomly into S disjoint folds, collecting the corresponding indices in sets denoted I_s, for $s \in [S]$. For $s \in [S]$, we define the complement of fold I_s as $I_s^c = [n]\backslash I_s$; and we let $s(i)$ denote the index of the fold containing observation i, so that $i \in s(i)$. Then $\widehat{\mu}_s(z) = \widehat{\mu}(z; \{W_i\}_{i\in I_s^c})$ is the estimator of $\mu_0(z)$ that uses all data points except those in the sth fold. Our proposed covariate-adjusted cross-fitting estimator of τ is then given by

$$\widehat{\tau}_{CF}(h; \widehat{\mu}) = \sum_{i=1}^{n} w_i(h) M_i(\widehat{\mu}_{s(i)}), \tag{3.6}$$

where $M_i(\widehat{\mu}_{s(i)}) = Y_i - \widehat{\mu}_{s(i)}(Z_i)$ is the covariate-adjusted outcome variable generated from the first-stage estimate of μ_0 based on data from the folds that do not contain the ith observation. In practice, $S = 5$ or $S = 10$ are common choices for the number of folds.

3.5. **Examples of Covariate Adjustments.**

A wide range of methods can be used in our framework to estimate the adjustment function μ_0. As mentioned in the overview, such estimates do not have to be consistent for μ_0 in order for $\widehat{\tau}_{CF}(h; \widehat{\mu})$ to be consistent for τ, making correct specification a more minor concern. This is essentially a consequence of the fact that (3.3) holds for *any* adjustment function, not just the optimal one. For efficiency, however, it is of course desirable to estimate μ_0 as accurately as possible. We generally focus on estimates of the form

$$\widehat{\mu}_s(z) = (\widehat{\mu}_s^+(z) + \widehat{\mu}_s^-(z))/2,$$

with $\widehat{\mu}_s^+(z)$ and $\widehat{\mu}_s^-(z)$ being estimates of $\mu_0^+(z) = \mathbb{E}[Y_i|X_i = 0^+, Z_i = z]$ and $\mu_0^-(z) = \mathbb{E}[Y_i|X_i = 0^-, Z_i = z]$, respectively. The following examples mention general types of procedures by which the estimates of the latter two quantities can be obtained.

Example 1 (Machine Learning)**.** Machine learning methods, such as lasso or post-lasso regression, random forests, deep neural networks, boosting, or ensemble combinations thereof, are well-suited for estimating conditional expectation functions of the form $\mathbb{E}[Y_i|X_i = x, Z_i = z]$ if they fall into the class of functions that such methods are designed to approximate. Such estimates are not naturally guaranteed to be very accurate for a particular value of x. However, such methods can be adapted to this task, for example, by restricting the respective algorithm to a subset of the data that only contains units whose realization of the running variable falls within some window on either side of the cutoff. To more formally describe this, let $\widehat{\mathbb{E}}[Y_i|Z_i = z; \{W_i : i \in [n]\}]$ be a generic machine learning estimator of $\mathbb{E}[Y_i|Z_i = z]$ computed on the full data, and let $b > 0$ be some positive bandwidth. Then we can put $\widehat{\mu}_s^+(z) = \widehat{\mathbb{E}}[Y_i|Z_i = z; \{W_i : i \in I_s^c, X_i \in (0, b)\}]$, and $\widehat{\mu}_s^-(z) = \widehat{\mathbb{E}}[Y_i|Z_i = z; \{W_i : i \in I_s^c, X_i \in (-b, 0)\}]$. The choice of b involves a bias-variance trade-off similar to the one encountered in classical nonparametric kernel regression problems.[4]

Example 2 (Classical Nonparametric Regression)**.** Under appropriate smoothness conditions, one can also use classical nonparametric regression techniques to estimate μ_0^+ and μ_0^-, with local polynomial regression being particularly suitable due to their good boundary properties. Series estimators are another possibility, but they can exhibit erratic behavior near the boundary in RD-type problems (Gelman and Imbens, 2019). One can address such issues, which appear analogously for other global smoothers, using the strategy described in the previous example.

Example 3 (Ordinary Least Squares)**.** In principle, one can also obtain estimates of μ_0 by specifying a simple parametric model for $\mathbb{E}[Y_i|X_i = x, Z_i = z]$, such as linear regression model with interactions that imply separate fits for units above and below the cutoff. We consider such estimators below mostly to illustrate that under the shape restriction they imply one can substantially weaken some of the assumptions used in our theoretical analysis, but some practitioners might find them attractive due to their simplicity.

3.6. **Relationship to Calonico et al. (2019).** As mentioned above, Calonico et al. (2019) study an estimator of τ based on including covariates linearly and without localization into the regression (2.2):

$$\widehat{\tau}_{CCFT}(h) = e_1^\top \operatorname*{argmin}_{\beta,\gamma} \sum_{i=1}^{n} K_h(X_i)(Y_i - (T_i, X_i, T_i X_i, 1)\beta - Z_i^\top \gamma)^2. \tag{3.7}$$

[4] For machine learning methods that are able to handle weighted data, such a restriction of the sample corresponds to assigning kernel weights based on a uniform kernel. With such methods, one could also use any other kernel weighting scheme, such as one based on the triangular kernel.

This estimator can be interpreted as a special case of our procedure that does not use cross-fitting. Specifically, let $\widehat{\gamma}_h$ denote the minimizer with respect to γ in (3.7), and put $\widehat{\mu}^+_{CCFT}(Z_i) = \widehat{\mu}^-_{CCFT}(Z_i) = Z_i^\top \widehat{\gamma}_h$, so that an estimate of μ_0 is given by $\widehat{\mu}_{CCFT}(Z_i) = Z_i^\top \widehat{\gamma}_h$. Then, by simple least squares algebra, we see that

$$\widehat{\tau}_{CCFT}(h) = \sum_{i=1}^{n} w_i(h) M_i(\widehat{\mu}_{CCFT})$$

fits into our general framework. The estimator $\widehat{\tau}_{CCFT}(h)$ can thus be interpreted as a version of our procedure that implicitly imposes the (generally incorrect) specification that $\mu_0^+(z) = \mu_0^-(z) = c + z^\top \gamma_0$ for some vector of coefficients γ_0 and some constant c. We stress again that correct specification of μ_0^+ and μ_0^- is not required for such an estimator of τ to be consistent.

4. MAIN RESULTS

In this section, we study the theoretical properties of our proposed estimators.

4.1. **Assumptions.** Most of the conditions we impose are either standard in the RD literature, or concern the general properties of the first-stage estimator $\widehat{\mu}$. To describe them, we first introduce some notation. We denote the support of Z_i by $\mathcal{Z}$, and the support of X_i by $\mathcal{X}$. We also let $\mathcal{X}_h = \mathcal{X} \cap [-h, h]$ and $\mathcal{Z}_h = \operatorname{supp}(Z_i | X_i \in \mathcal{X}_h)$. We also define the following class of functions that satisfy a suitable integrability condition:

$$\mathcal{M} = \Big\{ \mu : \mathcal{Z} \to \mathbb{R} \text{ s.t. } \sup_{x \in \mathcal{X}} \mathbb{E}[\mu(Z_i)^{2+\delta} | X_i = x] < \infty \text{ for some } \delta > 0 \Big\}.$$

Assumption 1. *The distribution of Z_i given $X_i = x$ converges setwise to the same limit as x tends to zero from either above or below. That is, $\mathbb{P}[Z_i \in B | X_i = 0^+] = \mathbb{P}[Z_i \in B | X_i = 0^-] \equiv \mathbb{P}[Z_i \in B | X_i = 0]$ for all Borel sets B.*

Assumption 1 formalizes the notion that the conditional distribution of the covariates does not change "much" with the value of the running variable around the cutoff. This matches the idea that the covariates are determined prior to the treatment assignment. The key implications of Assumption 1 for our analysis are that $\mathbb{E}[\mu(Z_i) | X_i = x]$ and $\mathbb{E}[\mu(Z_i)^2 | X_i = x]$ are continuous at $x = 0$ for all adjustment functions $\mu \in \mathcal{M}$.[5]

[5] If one imposes further restrictions on the elements of $\mathcal{M}$, continuity of $\mathbb{E}[\mu(Z_i)|X_i = x]$ and $\mathbb{E}[\mu(Z_i)^2|X_i = x]$ also follows under weaker conditions on the conditional distribution of Z_i given X_i. If $\mathcal{M}$ contained only continuous functions, for example, then setwise convergence in Assumption 1 could be replaced with convergence in distribution. If $\mathcal{M}$ contained only linear functions, then continuity of the first two conditional moments of $Z_i | X_i = x$ would be sufficient for obtaining our results.

Assumption 2. *(i) X_i is continuously distributed with density f_X, which is continuous and bounded away from zero over an open neighborhood of the cutoff; (ii) The kernel function K is a bounded and symmetric density function that is continuous on its support, and equal to zero outside some compact set, say $[-1, 1]$; (iii) The bandwidth satisfies $h \to 0$ and $nh \to \infty$ as $n \to \infty$.*

Assumption 2 collects some standard conditions from the RD literature. Note that continuity of the running variable's density f_X around the cutoff is strictly speaking not required for an RD analysis. However, a discontinuity in f_X is typically considered to be an indication of a design failure that prevents τ from being interpreted as a causal parameter (McCrary, 2008; Gerard et al., 2020). For this reason, we focus on the case of a continuous running variable density in this paper.

Assumption 3. *For all $n \in \mathbb{N}$, there exist a set $\mathcal{T}_n \subset \mathcal{M}$ and a function $\bar{\mu} \in \mathcal{T}_n$ such that: (i) $\widehat{\mu}_s$ belongs to $\mathcal{T}_n$ with probability approaching 1 for all $s \in [S]$; (ii) it holds that:*

$$\sup_{\mu \in \mathcal{T}_n} \sup_{x \in \mathcal{X}_h} \mathbb{E}\left[(\mu(Z_i) - \bar{\mu}(Z_i))^2 | X_i = x\right] = O(r_n^2)$$

for some deterministic sequence $r_n = o(1)$.

Assumption 3 states that the first-stage estimator belongs with high probability to some realization set $\mathcal{T}_n \subset \mathcal{M}$ that contracts around a deterministic sequence of functions in an L_2 sense (the supremum is taken over $\mathcal{X}_h$ instead of $\mathcal{X}$ as the properties of the first stage estimator are only relevant for observations that are used in the second-stage local linear regression). This assumption is rather weak, as it does not impose any conditions on the speed at which $\widehat{\mu}$ concentrates around $\bar{\mu}$, and also allows the function $\bar{\mu}$ to be different from the target function μ_0. The latter aspects allows, for example, to base $\widehat{\mu}$ on a specified parametric model of μ_0.

Mean-square error consistency as prescribed in Assumption 3 follows under classical conditions for the parametric and nonparametric procedures mentioned in Examples 2 and 3 above for settings in which the dimension of Z is fixed. For the type of implementation of machine learning estimators of μ_0 described in Example 1, existing results imply that for fixed $b > 0$ it holds that

$$\sup_{\mu \in \mathcal{T}_n} \mathbb{E}\left[(\mu(Z_i) - \bar{\mu}(Z_i))^2 | X_i \in (-b, b)\right] = O(r_n^2), \tag{4.1}$$

with $\bar{\mu}(z) = (\mathbb{E}[Y_i | X_i \in (-b, 0), Z_i = z] + \mathbb{E}[Y_i | X_i \in (0, b), Z_i = z])/2$ and some $r_n = o(1)$.

For example, if $\bar{\mu}(z)$ is contained in a Hölder class of order s, then (4.1) can hold with $r_n^2 = n^{-2s/(2s+d)}$ for estimators that exploit smoothness. If $\bar{\mu}(z)$ is s-sparse, then (4.1) can hold with $r_n^2 = n^{-s\log(d)/n}$ for estimators that exploit sparsity. Assumption 3 follows from (4.1) if the conditional distribution of the covariates does not change "too quickly" when moving away from the cutoff. For example, if the covariates are continuously distributed conditional on the running variable, having that

$$\sup_{x\in\mathcal{X}_h}\sup_{z\in\mathcal{Z}_h}\frac{f_{Z|X}(z|x)}{f_{Z|X\in(-b,b)}(z)} < C,$$

for some constant C and all n sufficiently large, suffices. Similar conditions can be given for discrete conditional covariate distributions, or intermediate cases. If $\mathbb{E}[Y_i|X_i = x, Z_i = z]$ is sufficiently smooth in x on both sides of the cutoff, we can also expect that $\bar{\mu}$ is "close" to μ_0 for small values of b.

Assumption 4. *For all $n \in \mathbb{N}$ and $j \in \{1, 2\}$, it holds that:*

$$\sup_{\mu\in\mathcal{T}_n}\sup_{x\in\mathcal{X}_h\setminus\{0\}}\left|\partial_x^j\mathbb{E}\left[\mu(Z_i)-\bar{\mu}(Z_i)|X_i = x\right]\right| = O(r_n).$$

for $r_n = o(1)$ as in Assumption 3.

Assumption 4 also concerns the first-stage estimator, and requires the first and second derivatives of $\mathbb{E}\left[\mu(Z_i)-\bar{\mu}(Z_i)|X_i = x\right]$ to be close to zero in large samples for all $\mu \in \mathcal{T}_n$. This can easily be verified if the functions contained in $\mathcal{T}_n$ have a sufficiently simple structure. For example, if $\mu(z)$ is linear for all $\mu \in \mathcal{T}_n$, then Assumption 4 holds if each component of $\mathbb{E}[Z_i|X_i = x]$ is twice continuously differentiable on $\mathcal{X} \setminus \{0\}$. Without imposing any structure on the functions contained in $\mathcal{T}_n$, Assumption 4 also follows from Assumption 3 and restrictions on the conditional distribution of Z_i given X_i. To give an example, suppose that (i) Assumption 3 holds; (ii) conditions allowing an application of dominated convergence are satisfied, so that

$$\partial_x^j\mathbb{E}[\mu(Z_i)-\bar{\mu}(Z_i)|X_i = x] = \int(\mu(z)-\bar{\mu}(z))\partial_x^j f_{Z|X}(z|x)dz; \tag{4.2}$$

and (iii) that

$$\sup_{x\in\mathcal{X}_h\setminus\{0\}}\mathbb{E}\left[\left(\frac{\partial_x^j f_{Z|X}(Z_i|x)}{f_{Z|X}(Z_i|x)}\right)^2|X_i = x\right] < C \tag{4.3}$$

for some constant C; for all n that are sufficiently large and for $j \in \{1, 2\}$. Then it follows from simple algebra that Assumption 4 holds. The integrability condition (4.3) itself holds, for example, if the conditional density $f_{Z|X}(z|x)$ is bounded away from zero and $\partial_x^j f_{Z|X}(z|x)$ is bounded for $j \in \{1,2\}$ uniformly in x and z. Simple calculations show that it also holds in a normal location model with twice continuously differentiable conditional expectation function. We note that it is straightforward to allow for different convergence rates in Assumptions 3 and 4 at the cost of slightly more involved notation. However, based on the above reasoning, we expect these rates to be the same in all relevant settings.

Assumption 5. *There exist constants C and L such that the following conditions hold for all $n \in \mathbb{N}$. (i) $\mathbb{E}[M_i(\bar{\mu})|X_i = x]$ is twice continuously differentiable on $\mathcal{X} \setminus \{0\}$ with L-Lipschitz continuous second derivative bounded by C; (ii) For all $x \in \mathcal{X}$ and some $q > 2$ $\mathbb{E}[(M_i(\bar{\mu}) - \mathbb{E}[M_i(\bar{\mu})|X_i])^q|X_i = x]$ exists and is bounded by C; (iii) $\mathbb{V}[M_i(\bar{\mu})|X_i = x]$ is L-Lipschitz continuous and bounded from below by $1/C$ for all $x \in \mathcal{X} \setminus \{0\}$.*

Assumption 5 is standard for an RD analysis with $M_i(\bar{\mu})$ as the outcome variable. Part (i) is satisfied if $\mathbb{E}[Y_i|X_i = x]$ and $\mathbb{E}[\bar{\mu}(Z_i)|X_i = x]$ are twice continuously differentiable to the left and to the right of the cutoff, but it does *not* require continuity of the derivatives of $\mathbb{E}[\bar{\mu}(Z_i)|X_i = x]$ *at* the cutoff. This assumption is thus analogous to one in Calonico et al. (2019), who impose that $\mathbb{E}[Z_i|X_i = x]$ is thrice continuously differentiable to the left and to the right of the cutoff but not necessarily at the cutoff. In our theoretical analysis below, imposing full continuity of $\partial_x^2 \mathbb{E}[\bar{\mu}(Z_i)|X_i = x]$ around the cutoff would simplify the formula for the asymptotic bias of our estimator. Parts (ii) and (iii) of Assumption 5 impose standard restrictions on conditional moments of the outcome variable.

4.2. **Main Results.** In this section, we study the asymptotic properties of our estimator. We define the following kernel constants: $\bar{\nu} = (\bar{\nu}_2^2 - \bar{\nu}_1\bar{\nu}_3)/(\bar{\nu}_2\bar{\nu}_0 - \bar{\nu}_1^2)$ and $\bar{\kappa} = \int_0^\infty (k(v)(\bar{\nu}_1 v - \bar{\nu}_2))^2 dv/(\bar{\nu}_2\bar{\nu}_0 - \bar{\nu}_1^2)^2$, where $\bar{\nu}_j = \int_0^\infty v^j k(v)dv$.

Theorem 1. *(i) Suppose that Assumptions 1–4 hold. Then*

$$\widehat{\tau}_{CF}(h; \widehat{\mu}) = \widehat{\tau}(h; \bar{\mu}) + O_p(r_n(h^2 + (nh)^{-1/2})).$$

(ii) Suppose that additionally Assumption 5 holds. Then

$$\sqrt{nh}\, V(\bar{\mu})^{-1/2} \left(\widehat{\tau}_{CF}(h; \widehat{\mu}) - \tau - B(\bar{\mu})h^2\right) \to \mathcal{N}(0,1),$$

where

$$\begin{aligned} B(\bar{\mu}) &= \frac{\bar{\nu}}{2}\left(\partial_x^2 \mathbb{E}[M_i(\bar{\mu})|X_i = x]\big|_{x=0^+} - \partial_x^2 \mathbb{E}[M_i(\bar{\mu})|X_i = x]\big|_{x=0^-}\right) + o(1), \\ V(\bar{\mu}) &= \frac{\bar{\kappa}}{f_X(0)}\left(\mathbb{V}[M_i(\bar{\mu})|X_i = 0^+] + \mathbb{V}[M_i(\bar{\mu})|X_i = 0^-]\right). \end{aligned}$$

(iii) For any two functions $\mu^{(a)}, \mu^{(b)} \in \mathcal{M}$, it holds that $V(\mu^{(a)}) < V(\mu^{(b)})$ if and only if $\mathbb{V}[\mu_0(Z_i) - \mu^{(a)}(Z_i)|X_i = 0] < \mathbb{V}[\mu_0(Z_i) - \mu^{(b)}(Z_i)|X_i = 0]$

Part (i) of Theorem 1 is our key technical result. It shows that our proposed estimator is asymptotically equivalent to its infeasible analogue that replaces the estimator $\widehat{\mu}$ with the deterministic sequence $\bar{\mu}$. The accuracy of this approximation increases with the rate at which r_n tends to zero, i.e., with the rate at which $\widehat{\mu}$ concentrates around $\bar{\mu}$. We emphasize however that this first-order asymptotic equivalence holds even if the first-stage estimator converges arbitrarily slowly. The estimator $\widehat{\tau}_{CF}(h; \widehat{\mu})$ is this insensitive to sampling variation in $\widehat{\mu}$ because it is a sample analogue of the moment function (3.3), which states that

$$\tau = \mathbb{E}[M_i(\mu)|X_i = 0^+] - \mathbb{E}[M_i(\mu)|X_i = 0^-],$$

and this moment function is fully insensitive to variation in μ over $\mathcal{M}$. Moment functions with a local form of insensitivity with respect to a nuisance function, so called Neyman orthogonality, are used extensively in the recent literature on two-stage estimators that use machine learning in the first stage (e.g. Belloni et al., 2017; Chernozhukov et al., 2018). Our RD setup is specially because it gives rise to a moment function with the above-stated global insensitivity, which is much stronger and allows for example estimates of the nuisance function μ_0 to be potentially inconsistent (as discussed in Section 5.2, a similarly globally insensitive moment function exists, for example, in certain randomized experiments).

Part (i) of Theorem 1 also suggests that we can use existing methods for bandwidth choice and inference in our setup by applying them to the generated data set $\{(X_i, M_i(\widehat{\mu}_{s(i)})) : i \in [n]\}$, and ignoring sampling uncertainty originating from $\widehat{\mu}$. For instance, one could use the method in Imbens and Kalyanaraman (2012) for bandwidth choice, or the methods in Calonico et al. (2014) or Armstrong and Kolesár (2020) to construct confidence intervals for τ. Such methods should retain their general theoretical properties under a combination of their original and our assumptions. We give a specific result for standard error estimation in Appendix B.

Part (ii) of Theorem 1 shows that our estimator is asymptotically normal, and gives explicit expressions for its asymptotic bias and variance. We remark that the bias expression simplifies substantially if we strengthen Assumption 5(i) and impose that $\partial_x^2 \mathbb{E}[\bar{\mu}(Z_i)|X_i = x]$ is continuous not only on each side of the cutoff, but also at the cutoff (for all $n \in \mathbb{N}$). In this case, we have that

$$B(\bar{\mu}) = \frac{\bar{\nu}}{2} \left(\partial_x^2 \mathbb{E}[Y_i|X_i = x]|_{x=0^+} - \partial_x^2 \mathbb{E}[Y_i|X_i = x]|_{x=0^-} \right) + o(1),$$

whose leading term is identical to that of the "no covariates" RD estimator's bias, and does not depend on the function $\bar{\mu}$.

Part (iii) of Theorem 1 gives conditions for one adjustment function to yield a smaller asymptotic variance than another. It implies in particular $V(\mu) \geq V(\mu_0)$ for all functions $\mu \in \mathcal{M}$, and thus that our estimator achieves its minimum asymptotic variance if $\bar{\mu} = \mu_0$. It also shows that even if $\bar{\mu} \neq \mu_0$, our proposed covariate adjustments still yield efficiency gains relative to a standard RD estimator that ignores the covariates in a very wide range of settings. In particular, denoting the asymptotic variance of the "no covariates" estimator by $V(0)$, we have that $V(\bar{\mu}) < V(0)$ if and only if $\mathbb{V}[\mu_0(Z_i) - \bar{\mu}(Z_i)|X_i = 0] < \mathbb{V}[\mu_0(Z_i)|X_i = 0]$, i.e. whenever $\bar{\mu}(Z_i)$ has some explanatory power for $\mu_0(Z_i)$ among units near the cutoff.

5. FURTHER RESULTS AND DISCUSSIONS

In this section, we discuss a variation of cross-fitting and an extension to fuzzy RD designs.

5.1. **Other Implementation of Cross-Fitting.** When cross-fitting is used in other context, a popular type of implementation is to create an overall estimate of the parameter of interest by averaging separate estimates from each data fold. In our context, such an estimator would be given by

$$\widetilde{\tau}_{CF}(h, \widehat{\mu}) = \frac{1}{S} \sum_{s \in [S]} \sum_{i \in I_s} w_{i,s}(h) M_i(\widehat{\mu}_s),$$

where $w_{i,s}(h)$ is the local linear regression weight of unit i based on the s-th fold of the data, which depends on the data through the realizations of the running variable only; see Appendix A.1 for an explicit expression (one could also have different bandwidths in each fold of the data). It follows from the proof of Theorem 1 that this estimator is asymptotically equivalent to our proposed procedure. That is,

$$\widetilde{\tau}_{CF}(h, \widehat{\mu}) - \widehat{\tau}_{CF}(h, \widehat{\mu}) = O_P(r_n(h^2 + (nh)^{-1/2})).$$

An analogous point is made by Chernozhukov et al. (2018) in the context of the (unconditional) average treatment effect estimation; cf. their methods DML1 and DML2. By inspecting the proof of Theorem 1, we can see that the expansion of $\widehat{\tau}_{CF}(h,\widehat{\mu}) - \widehat{\tau}(h,\bar{\mu})$ contains an extra bias term relative to an expansion of $\widetilde{\tau}_{CF}(h,\widehat{\mu}) - \widehat{\tau}(h,\bar{\mu})$. However, this term is of smaller order than $O_P(r_n(h^2 + (nh)^{-1/2}))$ and hence it does not affect our first-order asymptotic equivalence result. We prefer our proposed implementation because despite the small additional bias because it allows existing routines for bandwidth selection and confidence interval construction to be applied directly to the modified data $\{(M_i(\widehat{\mu}_{s(i)}), X_i)\}_{i\in[n]}$.

5.2. **Analogies with Randomized Experiments.** The results in Section 4 are qualitatively very similar to ones that have been obtained for efficient influence function (EIF) estimators of the population average treatment effect (PATE) in randomized experiments with constant propensity scores (e.g., Wager et al., 2016; Chernozhukov et al., 2018). Such parallels arise because our covariate-adjusted RD estimator is in many ways a direct analogue of such EIF estimators. To see this, consider the following sketch of the latter's properties.

Consider a randomized experiment with unconfounded treatment assignment and constant propensity score p. Using our notation in an analogous fashion, the EIF of the PATE in such a setup is typically given in the literature (e.g., Hahn, 1998) in the form

$$\psi_i(m_0^0, m_0^1) = m_0^1(Z_i) - m_0^0(Z_i) + \frac{T_i(Y_i - m_0^1(Z_i))}{p} - \frac{(1-T_i)(Y_i - m_0^0(Z_i))}{1-p},$$

where $m_0^t(z) = \mathbb{E}[Y_i|Z_i = z, T_i = t]$ for $t \in \{0,1\}$. The minimum variance any regular estimator of the PATE can achieve is thus given by $V_{\text{PATE}} = \mathbb{V}(\psi_i(m_0^0, m_0^1))$. By randomization, it also holds $\tau_{\text{PATE}} = \mathbb{E}[\psi_i(m^0, m^1)]$ for all (suitably integrable) functions m^0 and m^1, and thus the PATE is identified by a moment function that satisfies a global invariance property. A sample analogue estimator of τ_{PATE} based on this moment function reaches has asymptotic variance V_{PATE} if $\widehat{m}^t$ is a consistent estimator of m_0^t for $t \in \{0,1\}$, but remains consistent and asymptotically normal with asymptotic variance $\mathbb{V}(\psi_i(\bar{m}^0, \bar{m}^1))$ if $\widehat{m}^t$ is consistent for some other function $\bar{m}^t$, $t \in \{0,1\}$. The convergence of $\widehat{m}^t$ to $\bar{m}^t$ can be arbitrarily slow for these results (e.g. Wager et al., 2016; Chernozhukov et al., 2018). These findings are clearly similar to ours in Section 4.

To see the analogy to our RD setup, note that if we write $m(z) = (1-p)m^1(z) + p\,m^0(z)$ for any two functions m^0 and m^1, so that $m_0(z) = (1-p)m_0^1(z) + p\,m_0^0(z)$, the PATE's influence function can also be expressed as

$$\psi_i(m_0^0, m_0^1) = \frac{T_i(Y_i - m_0(Z_i))}{p} - \frac{(1-T_i)(Y_i - m_0(Z_i))}{1-p}.$$

The globally insensitive moment function $\mathbb{E}[\psi_i(m^0, m^1)]$ on which the EIF estimat is based can thus be written as

$$\mathbb{E}[\psi_i(m^0, m^1)] = \mathbb{E}[Y_i - m(Z_i)|T_i = 1] - \mathbb{E}[Y_i - m(Z_i)|T_i = 0],$$

which is the difference in average covariate-adjusted outcomes between treated and untreated units. This is fully analogous to our equation (3.3), with $p = 1/2$, and conditioning on $T_i = 1$ and $T_i = 0$ replaced by conditioning on X_i in infinitesimal right and left neighborhoods of the cutoff. The value $p = 1/2$ is appropriate here because continuity of the running variable's density implies that an equal share of units close to the cutoff can be found on either side. An EIF estimator of τ_{PATE} is thus analogous to our estimator $\widehat{\tau}_{CF}(h; \widehat{\mu})$, as they are both sample analogues a moment function with the same basic properties. We also note that with $p = 1/2$ it holds that

$$V_{\text{PATE}} = 2 \times (\mathbb{V}[Y_i - m_0(Z_i)|T_i = 1] + \mathbb{V}[Y_i - m_0(Z_i)|T_i = 0]),$$

which is analogous to our formula for $V(\mu_0)$ in Section 4.

5.3. **Fuzzy RD Designs.** In fuzzy RD designs, units are assigned to treatment if their realization of the running variable falls above the threshold value, but they do not necessarily comply with this assignment. The conditional treatment probability hence jumps at the cutoff, but in contrast to sharp RD designs it generally does not jump from zero to one. The parameter of interest in fuzzy RD designs is

$$\theta = \frac{\tau_Y}{\tau_T} \equiv \frac{\mathbb{E}[Y_i|X_i = 0^+] - \mathbb{E}[Y_i|X_i = 0^-]}{\mathbb{E}[T_i|X_i = 0^+] - \mathbb{E}[T_i|X_i = 0^-]},$$

which is the ratio of two sharp RD estimands. Under standard conditions (Hahn et al., 2001; Dong, 2017), one can interpret θ as the average causal effect of the treatment among units at the cutoff whose treatment decision is affected by whether their value of the running variable is above or below the cutoff.

Building on our proposed method, we can estimate θ by the ratio of two sharp, covariate-adjusted RD estimators:

$$\widehat{\theta}_{CF}(h;\widehat{\mu}_Y,\widehat{\mu}_T)=\frac{\widehat{\tau}_{CF,Y}(h,\widehat{\mu}_Y)}{\widehat{\tau}_{CF,T}(h,\widehat{\mu}_T)}=\frac{\sum_{i=1}^n w_i(h)(Y_i-\widehat{\mu}_{Y,s(i)}(Z_i))}{\sum_{i=1}^n w_i(h)(T_i-\widehat{\mu}_{T,s(i)}(Z_i))},$$

where the notation is analogous to that used before, with the subscripts Y and T indicating the respective outcome variable. That is, $\widehat{\mu}_Y$ denotes an estimate of $\mu_{Y,0}(z)=(\mathbb{E}[Y_i|X_i=0^+,Z_i=z]+\mathbb{E}[Y_i|X_i=0^-,Z_i=z])/2$, and $\widehat{\mu}_T$ is an estimate of $\mu_{T,0}(z)=(\mathbb{E}[T_i|X_i=0^+,Z_i=z]+\mathbb{E}[T_i|X_i=0^-,Z_i=z])/2$, etc. If $|\tau_T|$ is bounded away from zero and Assumptions 1–5 hold with T_i replacing Y_i, it then follows from Theorem 1 and the Delta method that $\widehat{\theta}_{CF}(h;\widehat{\mu}_Y,\widehat{\mu}_T)-\theta$ is asymptotically equivalent to a sharp RD estimator with the infeasible outcome variable

$$U_i(\bar{\mu}_Y,\bar{\mu}_T)=\frac{1}{\tau_T}\left(Y_i-\theta T_i-(\bar{\mu}_Y(Z_i)-\theta\bar{\mu}_T(Z_i))\right),$$

which in turn implies that

$$\sqrt{nh}\,V_\theta(\bar{\mu}_Y,\bar{\mu}_T)^{-1/2}\left(\widehat{\theta}_{CF}(h;\widehat{\mu}_Y,\widehat{\mu}_T)-\theta-B_\theta(\bar{\mu}_Y,\bar{\mu}_T)h^2\right)\to\mathcal{N}(0,1),$$

where

$$\begin{aligned}B_\theta(\bar{\mu}_Y,\bar{\mu}_T)&=\frac{\bar{\nu}}{2}\left(\partial_x^2\mathbb{E}[U_i(\bar{\mu}_Y,\bar{\mu}_T)|X_i=x]\big|_{x=0^+}-\partial_x^2\mathbb{E}[U_i(\bar{\mu}_Y,\bar{\mu}_T)|X_i=x]\big|_{x=0^-}\right)+o(1),\\ V_\theta(\bar{\mu}_Y,\bar{\mu}_T)&=\frac{\bar{\kappa}}{f_X(0)}\left(\mathbb{V}[U_i(\bar{\mu}_Y,\bar{\mu}_T)|X_i=0^+]+\mathbb{V}[U_i(\bar{\mu}_Y,\bar{\mu}_T)|X_i=0^-]\right).\end{aligned}$$

It is also easily seen that $V_\theta(\mu_Y,\mu_T)$ is minimized by the function pair $(\mu_{Y,0},\mu_{T,0})$ defined above. That is, even though $\mu_{Y,0}$ and $\mu_{T,0}$ were obtained as separate minimizers of the asymptotic variance of estimates of τ_Y and τ_T, respectively, the pair turns out to also minimize the asymptotic variance of the ratio estimator of θ.

The above normality result could be used to construct a confidence interval for θ. However, given the issues with such "Delta method" inference, we recommend constructing confidence sets for θ via the Anderson-Rubin-type approach in Noack and Rothe (2021).

6. SIMULATIONS

In this section, we compare the finite sample performance of our proposed estimator for different first-stage estimation methods in a Monte Carlo study.

6.1. **Setup.** We consider four different data generating processes (DGPs) in our simulations, which are index by a parameter $L \in \{0, 4, 10, 25\}$ that determines the complexity of how the covariates can affect the outcome. In each DGP, the running variable X_i follows the uniform distribution over $[-1, 1]$, there are four independent covariates Z_i, which are distributed uniformly over $[-1 + x^2, 1 + x^2]^4$ conditional on $X_i = x$, and the outcome is generated as:

$$Y_i = \mathbf{1}\{X_i \geq 0\} + \mu_L(X_i, Z_i) + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, 0.25),$$

$$\mu_L(X_i, Z_i) = \operatorname{sign}(X_i) \cdot (X_i + X_i^2 - 2\,(X_i - 0.1)_+^2) + \bar{\iota}_L(\rho) \sum_{l=1}^{L} b_l(Z_i),$$

where the terms $b_l(Z_i)$, $l \in [L]$, are Hermite polynomials of the covariates. For positive L and ρ, we choose the coefficient $\bar{\iota}_L(\rho)$ such that $\mathbb{V}[\mu_L(0, Z_i)|X_i = 0] = \rho^2 \mathbb{V}[\varepsilon_i]$. In this definition, ρ represents the signal to noise ratio at the cutoff given the treatment status. It determines the scope for improvements from using covariates, but it does not affect the relative performance of different covariate adjustments. We report simulation results for $\rho = 3$ in the main text, and for further values in Appendix C. The results are based on $5,000$ replications of sample with size $n = 2,000$.

We consider in total seven implementations of our proposed procedure with different first-stage estimators. For baseline comparisons, we consider: (i) the standard RD estimator with no covariate adjustments; (ii) the infeasible, optimal RD estimator with covariate adjustments based on the true conditional expectation function; (iii) the infeasible RD estimator with adjustments based on the best linear prediction on the population level of the true conditional expectation function given the four baseline covariates. We also consider four feasible adjustment functions based on: (iv) a linear regression given the four covariates; (v) a local linear regression given the four covariates; (vi) a post-lasso regression given a total of 200 transformation terms based on the four covariates; and (vii) a random forest with the four baseline covariates. In the first stage of the feasible procedures, the observations are weighted using triangular kernel weights with the bandwidth that was used for the standard RD estimator without covariates.

In each simulation run, we use the bias-aware approach of Armstrong and Kolesár (2018, 2020) to select the bandwidth for our estimators, and to compute a confidence interval. This requires specifying a smoothness bound, which we set to its population value in each DGP. The main qualitative conclusions of our simulation study also hold with other methods for bandwidth selection and confidence interval construction, such as robust bias corrections and undersmoothing. We present these results in Appendix C. There we also compare our

Table 1: Simulation Results

	Cov	Bias	SD	CI	h	Cov	Bias	SD	CI	h
	DGP 1: L=0					**DGP 2: L=4**				
Standard	.970	-.014	.074	.324	.432	.961	-.071	.186	.818	.688
Optimal Inf	.970	-.014	.074	.324	.432	.966	-.015	.075	.325	.432
Linear Inf	.970	-.014	.074	324	432	966	-15	75	.325	.432
Linear	970	-.014	.074	.327	.433	.967	-.015	.075	.326	.433
Local Linear	.970	-.014	.074	.327	.433	.968	-.014	.075	.327	.433
Lasso	.967	-.014	.076	.331	.436	.966	-.021	.088	.383	.466
Forest	.968	-.015	.076	.331	.436	.967	-.021	.087	.379	.465
	DGP 3: L=10					**DGP 4: L=25**				
Standard	.964	-.095	.191	.876	.793	.959	-.063	.185	.810	.685
Optimal Inf	.965	-.013	.076	.325	.432	.969	-.013	.074	.324	.432
Linear Inf	.967	-.048	.127	.562	.618	.968	-.043	.103	.472	.590
Linear	.959	-.040	.137	.591	.597	.965	-.043	108	.492	588
Local Linear	.963	-.016	.083	.356	.452	.968	-.016	82	.359	.456
Lasso	.962	-.020	.092	.391	.467	.968	-.014	.077	.340	.443
Forest	.966	-.019	.085	.372	.469	.971	-.022	.093	.413	.490

Notes: Results based on 5000 Monte Carlo draws. Columns show results for simulated coverage of confidence intervals with 95% nominal level (Cov); the bias (Bias); the Standard Deviation (SD); the average confidence interval length (CI); and the average selected bandwidth (h).

estimators to the linear covariates adjustment method proposed by Calonico et al. (2019).[6]

6.2. **Simulations Results.** Table 1 reports estimation and inference results for different types of adjustments. All CIs have simulated coverage rates slightly above the nominal one. This is because the bias-aware approach accounts for the "worst-case" smoothing bias, which is not achieved in the considered models. First, we compare the standard RD estimator and the infeasible estimators. In DGP 1, these estimators are numerically equal. In DGPs 2–4, where the covariates have some explanatory power for the outcome, the infeasible estimators have a substantially lower standard deviation than the standard estimator has. If the linear

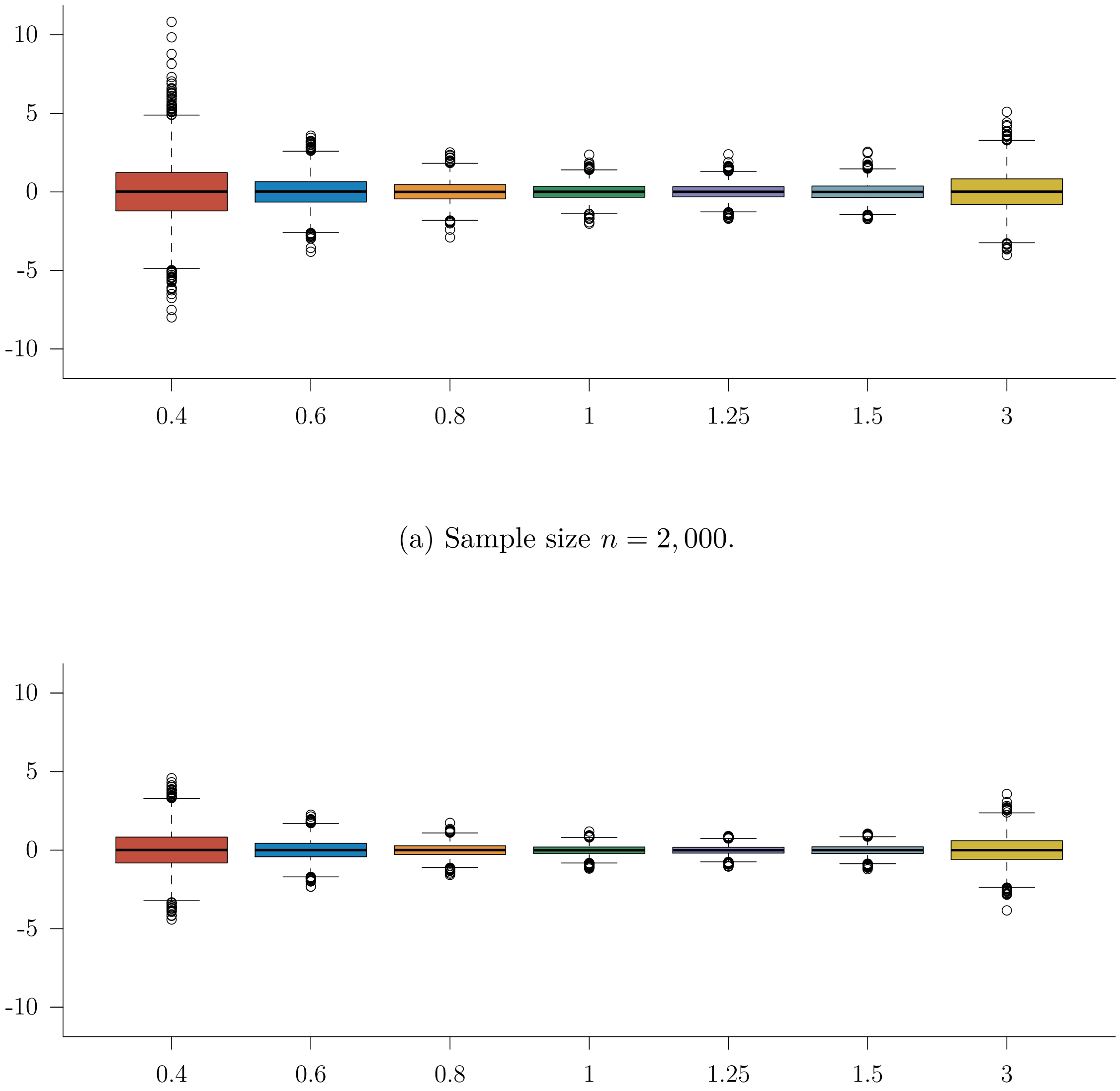
[6] All computations are carried out with the statistical software `R`. The Hermit-polynomials are computed using the package `calculus`. To implement the first-stage estimators, we use the following packages: `np` for local polynomial regressions; `glmnet` for lasso regressions; `grf` for random forests, where predictions are based on 200 trees. In the second stage, a triangular kernel is used and EHW standard errors are computed. The bias-aware approach is based on the package `RDHonest`, and the other two approaches are implemented using the package `rdrobust`.

model is misspecified, the standard deviation of the optimal infeasible estimator is much smaller than that of the infeasible estimator with linear adjustments. We now turn to the feasible covariate-adjusted RD estimators. As predicted by Theorem 1, their standard deviations are close to those of their respective infeasible counterpart, with only a very minor increase due to first-stage estimation error.

In Figures 1 and 2, we compare the difference between the optimal infeasible RD estimator and two feasible ones: namely those with adjustments based on local linear regression and post-lasso regression for several choices of the tuning parameters (in each simulation draw, we find the MSE-optimal tuning parameters via cross-validation, and then scale it down or up by different factors). To facilitate comparisons of different covariate adjustments, in each simulation draw we use the bandwidth selected for the standard RD estimator in the second stage across all different methods. We consider two sample sizes, $n = 2,000$ and $n = 10,000$. We normalize the difference by the standard error of the optimal infeasible RD estimator.

In Figure 1, we observe that the normalized difference between the estimators is relatively small for a wide range of bandwidths around the optimal one. By comparing panels (a) and (b), we can see that these normalized differences become smaller as the sample size increases, which illustrates the asymptotic equivalence result in part (i) of Theorem 1. For a given sample size, the average absolute value of the normalized differences is U-shaped as a function of the bandwidth. If the bandwidth chosen in the first stage is too small, then the local linear estimator is very unstable. In this case, the property in Assumption 3 is not a good description of its finite-sample behavior, and the equivalence result in Theorem 1 fails. If the bandwidth is chosen to be too large, the local linear estimator has a relatively small variance, but it might be heavily biased, and it is effectively very similar to the linear estimator. In this case, the equivalence to an infeasible estimator holds with a different limiting sequence $\bar{\mu}$. We expect the estimator to be less efficient, but we emphasize that our inference procedure remain valid in this case.

Figure 2 shows a very similar pattern as Figure 1. If the penalty parameter in the lasso regression is chosen to be too small, effectively all covariates are classified as relevant, and the first-stage estimator has a high variance. In contrast, if the penalty parameter is chosen to be too large, very few covariates are classified as relevant. In this case, the RD estimator behaves similarly to the standard RD estimator.

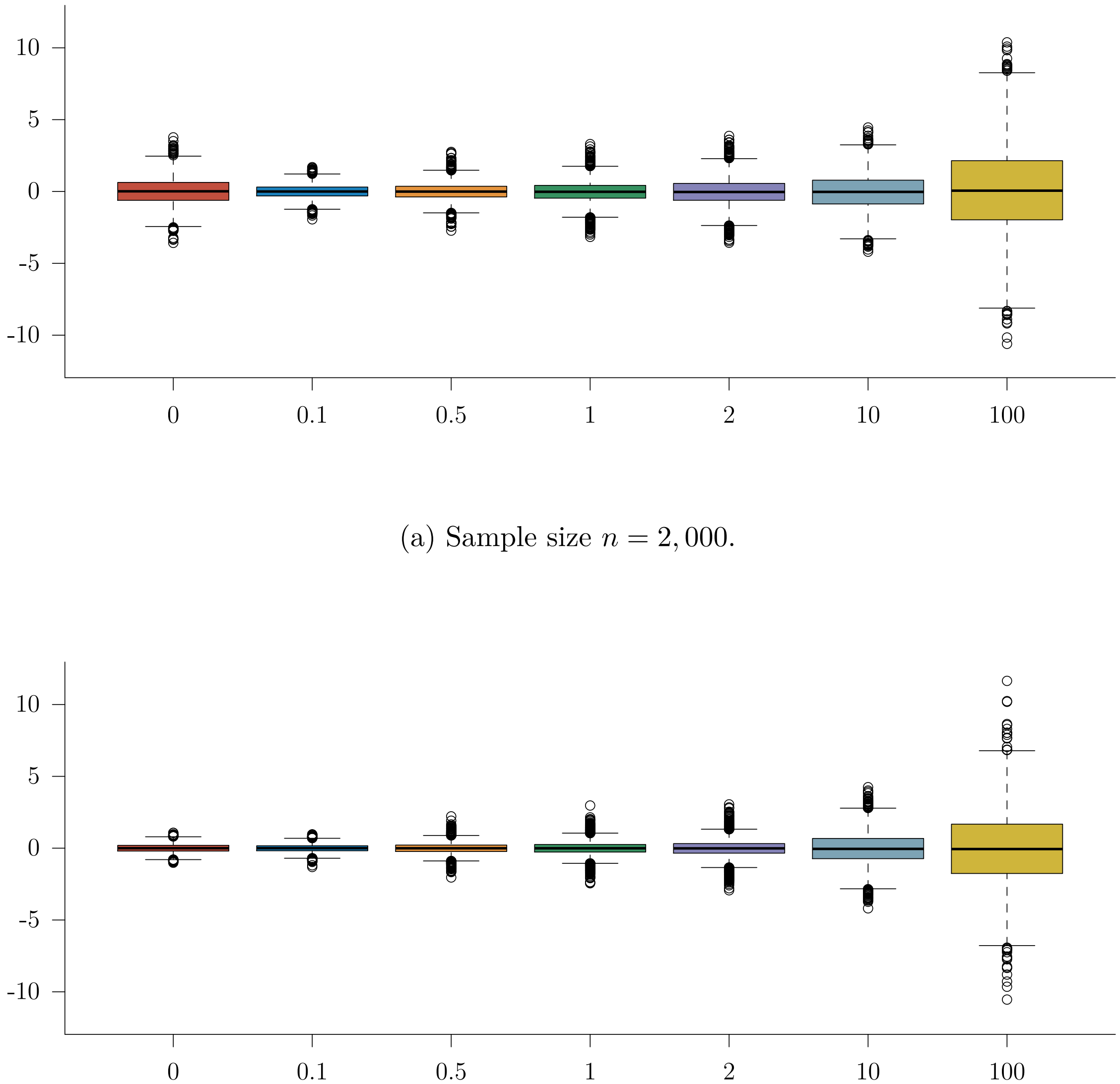


(a) Sample size $n = 2,000$.

(b) Sample size $n = 10,000$.

Figure 1: Normalized difference of RD estimates with local linear adjustments.

Notes: Difference between optimal infeasible and feasible RD estimate normalized by standard deviation of infeasible estimator. We consider various scaling factors for the cross-validated MSE-optimal first-stage bandwidth. Simulations are based on DGP 3. Panel (a) shows simulation results for $n = 2,000$, and Panel (b) for $n = 10,000$.



(a) Sample size $n = 2,000$.

(b) Sample size $n = 10,000$.

Figure 2: Normalized difference of RD estimates with post-lasso regression adjustments.
Notes: Difference between optimal infeasible and feasible RD estimate normalized by standard deviation of infeasible estimator. We consider various scaling factors for the cross-validated MSE-optimal first-stage penalty parameter. Simulations are based on Model 3. Panel (a) shows simulation results for $n = 2,000$, and Panel (b) for $n = 10,000$.

7. CONCLUSIONS

We have propose a novel class of estimators that can make use of covariate information more efficiently than the linear adjustment estimators that are currently used widely in practice. In particular, our approach allows the use of modern machine learning tools to adjust for covariates, and is at the same time largely unaffected by the "curse of dimensionality". Our estimator is also easy to implement in practice, and can be combined in a straightforward manner with existing methods for bandwidth choice and the construction of confidence intervals. For this reason, we expect it to be attractive for a wide range of economic applications.

A. PROOFS OF MAIN RESULTS

A.1. **Additional Notation.** We use the following notation throughout the proofs. The realizations of the running variable are denoted by $\mathbb{X}_n = \{X_i\}_{i\in[n]}$. For $s \in [S]$, $i \in I_{s(i)}$, and $j \in \{0,1\}$, we define the local linear weights as

$$\begin{aligned}
w_{i,s}^{(j)}(h) &= w_{i,s,+}^{(j)}(h) - w_{i,s,-}^{(j)}(h),\\
w_{i,s,+}^{(j)}(h) &= e_{j+1}^\top Q_{s,+}^{-1}\widetilde{X}_i K(X_i/h)\mathbf{1}\{X_i \geq 0\}, \quad Q_{s,+} = \sum_{i\in I_s} K(X_i/h)\widetilde{X}_i, \widetilde{X}_i^\top \mathbf{1}\{X_i \geq 0\},\\
w_{i,s,-}^{(j)}(h) &= e_{j+1}^\top Q_{s,-}^{-1}\widetilde{X}_i K(X_i/h)\mathbf{1}\{X_i < 0\}, \quad Q_{s,-} = \sum_{i\in I_s} K(X_i/h)\widetilde{X}_i\widetilde{X}_i^\top \mathbf{1}\{X_i < 0\},
\end{aligned}$$

with $\widetilde{X}_i = (1, X_i)^\top$. We omit the index s if the sum is taken over the whole sample and we omit the superscript (j) if $j = 0$. Further, for $\mu \in \mathcal{M}$, we let

$$\begin{aligned}
T_{s,+}(\mu) &= \sum_{i\in I_s} K(X_i/h)\widetilde{X}_i\mu(Z_i)\mathbf{1}\{X_i \geq 0\}\\
T_{s,-}(\mu) &= \sum_{i\in I_s} K(X_i/h)\widetilde{X}_i\mu(Z_i)\mathbf{1}\{X_i < 0\}.
\end{aligned}$$

With $m(x;\mu) = \mathbb{E}[\bar{\mu}(Z_i)-\mu(Z_i)|X_i = x]$, we also define $\beta_0(\mu) = m(0;\mu)$, $\beta_1^\star(\mu) = \partial_x m(x;\mu)|_{x=0^\star}$, and further $\beta^\star(\mu) = (\beta_0(\mu), \beta_1^\star(\mu))$ for $\star \in \{+,-\}$. Let $H = \text{diag}(1,h)$ and $\mathbb{I}_2 = \text{diag(1,1)}$.

A.2. **Proof of Theorem 1.** The proof of Theorem 1 is preceded by two lemmas.

Lemma A.1. *Suppose that Assumption 2 holds. Then for all $s \in [S]$ it holds that:*

(i) For all $j \in \mathbb{N}$,

$$\frac{1}{nh}\sum_{i=1}^{n} K(X_i/h)(X_i/h)^j T_i = \bar{\nu}_j f_X(0^+) + o_P(1),$$
$$\frac{1}{nh}\sum_{i=1}^{n} K(X_i/h)(X_i/h)^j (1-T_i) = \bar{\nu}_j f_X(0^-) + o_P(1),$$
$$\frac{1}{nh}\sum_{i=1}^{n} K(X_i/h)(X_i/h)^j T_i = \frac{S}{nh}\sum_{i\in I_s} K(X_i/h)(X_i/h)^j T_i + O_P((nh)^{-1/2}),$$
$$\frac{1}{nh}\sum_{i=1}^{n} K(X_i/h)(X_i/h)^j (1-T_i) = \frac{S}{nh}\sum_{i\in I_s} K(X_i/h)(X_i/h)^j (1-T_i) + O_P((nh)^{-1/2}).$$

(ii) For $j \in \{0,1\}$, $h^{2j}\sum_{i\in I_s} w_{i,s}^{(j)}(h)^2 = O_P((nh)^{-1})$ and $h^j \sum_{i\in I_s} |w_{i,s}^{(j)}(h) X_i^2| = O_P(h^2)$.

Proof. This follows from standard kernel calculations. □

Lemma A.2. *Suppose that Assumptions 1–4 hold. Then*

$$G_{s,\star}^{(j)} \equiv e_{j+1}^\top H(Q_{s,\star}^{-1} T_{s,\star}(\bar{\mu} - \widehat{\mu}_s) - \beta^\star(\bar{\mu} - \widehat{\mu}_s)) = O_p(r_n(h^2 + (nh)^{-1/2}))$$

for all $s \in [S]$, $\star \in \{+,-\}$, and $j \in \{0,1\}$.

Proof. We analyze the expectation and variance of $G_{s,\star}^{(j)}$ conditional on $\mathbb{X}_n$ and $(W_j)_{j\in I_s^c}$. First, we consider the expectation. It holds with probability approaching one that

$$\begin{aligned}|\mathbb{E}[G_{s,\star}^{(j)}|\mathbb{X}_n, (W_j)_{j\in I_s^c}]| &= \left|\sum_{i\in I_s} w_{i,s,\star}^{(j)}(h)\mathbb{E}[\bar{\mu}(Z_i) - \widehat{\mu}_s(Z_i)|X_i, (W_j)_{j\in I_s^c}]\right| \\ &\le \sup_{\mu\in\mathcal{T}_n}\left|\sum_{i\in I_s} w_{i,s,\star}^{(j)}(h)\mathbb{E}[\bar{\mu}(Z_i) - \mu(Z_i)|X_i]\right|\end{aligned}$$

By Taylor's theorem with the mean-value form of the remainder, it holds that

$$m(X_i;\mu) = m(0;\mu) + \partial_x m(x;\mu)|_{x=0^\star} X_i + \frac{1}{2}\partial_x^2 m(\widetilde{x}_i;\mu) X_i^2,$$

for some $\widetilde{x}_i$ between 0 and X_i. Using standard local linear algebra and the triangle inequality, we obtain that

$$\begin{aligned}
|\mathbb{E}[G_{s,\star}^{(j)}|\mathbb{X}_n, (W_j)_{j\in I_s^c}]| &\leq \sup_{\mu\in\mathcal{T}_n} \left| \frac{1}{2}\sum_{i\in I_s} w_{i,s,\star}^{(j)}(h)\partial_x^2 m(\widetilde{x}_i;\mu)X_i^2 \right| \\
&\leq \sup_{\mu\in\mathcal{T}_n} \sup_{x\in\mathcal{X}_h\setminus\{0\}} \frac{1}{2}|\partial_x^2 m(x;\mu)| \sum_{i\in I_s} \left| w_{i,s,\star}^{(j)}(h)X_i^2 \right| = O_p(r_n h^2),
\end{aligned}$$

where we use Lemma A.1 and Assumption 4 in the last step.

Second, we consider the conditional variance. It holds with probability approaching one that

$$\begin{aligned}
\mathbb{V}\left[G_{s,\star}^{(j)}|\mathbb{X}_n, (W_j)_{j\in I_s^c}\right] &= \sum_{i\in I_s} w_{i,s,\star}^{(j)}(h)^2 \mathbb{V}\left[\bar{\mu}(Z_i) - \widehat{\mu}_s(Z_i)|\mathbb{X}_n, (W_j)_{j\in I_s^c}\right] \\
&\leq \sup_{\mu\in\mathcal{T}_n} \sum_{i\in I_s} w_{i,s,\star}^{(j)}(h)^2 \mathbb{E}[(\bar{\mu}(Z_i) - \mu(Z_i))^2|X_i] \\
&\leq \sup_{\mu\in\mathcal{T}_n} \sup_{x\in\mathcal{X}_h} \mathbb{E}[(\bar{\mu}(Z_i) - \mu(Z_i))^2|X_i = x] \sum_{i\in I_s} w_{i,s,\star}^{(j)}(h)^2 \\
&= O_p({r_n}^2(nh)^{-1}),
\end{aligned}$$

where we use Lemma A.1 and Assumption 3 in the last step. The conditional convergence implies the unconditional one (see Chernozhukov et al., 2018, Lemma 6.1), which concludes the proof. □

Proof of Theorem 1. We prove the three parts separately.
Part (i) It holds that:

$$\begin{aligned}
&\widehat{\tau}_{CF}(h;\widehat{\mu}) - \widehat{\tau}(h;\bar{\mu}) \\
&= e_1^\top \sum_{s=1}^{S} \left\{ Q_+^{-1} T_{s,+}(\bar{\mu} - \widehat{\mu}_s) - Q_-^{-1} T_{s,-}(\bar{\mu} - \widehat{\mu}_s) \right\} \\
&= e_1^\top \sum_{s=1}^{S} Q_+^{-1} Q_{s,+}(Q_{s,+}^{-1} T_{s,+}(\bar{\mu} - \widehat{\mu}_s) - \beta^+(\bar{\mu} - \widehat{\mu}_s)) + e_1^\top \sum_{s=1}^{S} Q_+^{-1} Q_{s,+}\, \beta^+(\bar{\mu} - \widehat{\mu}_s) \\
&\quad - e_1^\top \sum_{s=1}^{S} Q_-^{-1} Q_{s,-}(Q_{-,s}^{-1} T_{s,-}(\bar{\mu} - \widehat{\mu}_s) - \beta^-(\bar{\mu} - \widehat{\mu}_s)) - e_1^\top \sum_{s=1}^{S} Q_-^{-1} Q_{s,-}\, \beta^-(\bar{\mu} - \widehat{\mu}_s). \\
&\equiv A_1 + A_2 - A_3 - A_4.
\end{aligned}$$

In the following, we consider each of the four terms separately. First, note that

$$A_1 = e_1^\top H^{-1} \sum_{s=1}^{S} HQ_+^{-1}HH^{-1}Q_{s,+}H^{-1}H(Q_{s,+}^{-1}T_{s,+}(\bar{\mu} - \widehat{\mu}_s) - \beta^+(\bar{\mu} - \widehat{\mu}_s))$$

By Lemma A.1, for all $s \in [S]$, it holds that

$$HQ_+^{-1}HH^{-1}Q_{s,+}H^{-1} = \frac{1}{S}\mathbb{I}_2 + O_P((nh)^{-1/2}), \tag{A.1}$$

where throughout the proof we assume that the term $O_P((nh)^{-1/2})$ has conformable dimensions. Using Lemma A.2 and noting that $e_1^\top H^{-1} = e_1^\top$, we obtain that $A_1 = O_p(r_n(h^2 + (nh)^{-1/2}))$.

Second, it holds that

$$A_2 = e_1^\top H^{-1} \sum_{s=1}^{S} HQ_+^{-1}HH^{-1}Q_{s,+}H^{-1}H\,\beta^+(\bar{\mu} - \widehat{\mu}_s).$$

Using equation (A.1), we obtain that

$$\begin{aligned} A_2 &= \frac{1}{S}\sum_{s=1}^{S}(e_1^\top + O_p((nh)^{-1/2}))H\,\beta^+(\bar{\mu} - \widehat{\mu}_s) \\ &= \frac{1}{S}\sum_{s=1}^{S}\beta_0(\bar{\mu} - \widehat{\mu}_s)(1 + O_p((nh)^{-1/2})) + h\beta_1^+(\bar{\mu} - \widehat{\mu}_s)O_p((nh)^{-1/2}) \\ &= \frac{1}{S}\sum_{s=1}^{S}\beta_0(\bar{\mu} - \widehat{\mu}_s) + O_p(r_n(nh)^{-1/2}), \end{aligned}$$

where we use the fact $\beta_0(\bar{\mu} - \widehat{\mu}_s) = O_p(r_n)$ by Assumption 3 and $h\beta_1^+(\bar{\mu} - \widehat{\mu}_s) = O_p(r_n h)$ by Assumption 4 for all $s \in [S]$.[7]

Using analogous calculations, we can show that $A_3 = O_P(r_n(h^2 + (nh)^{-1/2}))$ and $A_4 = \frac{1}{S}\sum_{s=1}^{S}\beta_0(\bar{\mu} - \widehat{\mu}_s) + O_p(r_n(nh)^{-1/2})$, which concludes the proof of part (i).

Part (ii). By the conditional version of Lyapunov CLT, we obtain that

$$\mathrm{se}(h;\bar{\mu})^{-1}(\widehat{\tau}(h;\bar{\mu}) - \mathbb{E}[\widehat{\tau}(h;\bar{\mu})|\mathbb{X}_n]) \to \mathcal{N}(0,1).$$

where $\mathrm{se}^2(h;\bar{\mu}) = \sum_{i=1}^{n} w_i(h)^2\mathbb{V}[M_i(\bar{\mu})|X_i]$. Using L-Lipschitz continuity of $\mathbb{V}[M_i(\bar{\mu})|X_i = x]$

[7] Theorem 1 remains true if we weaken our Assumption 4 for the first derivative ($j = 1$) to: $\sup_{\mu\in\mathcal{T}_n}\left|\partial_x^1\mathbb{E}\left[\mu(Z_i) - \bar{\mu}(Z_i)|X_i = x\right]|_{x=0^\star}\right| = O(r_n/h)$ for $\star \in \{+,-\}$. However, we expect that the convergence rates for the first and second derivative will typically be the same.

in x, we obtain that

$$\mathrm{se}^2(h;\bar{\mu}) = \sum_{i=1}^{n} w_{i,-}(h)^2 \mathbb{V}[M_i(\bar{\mu})|X_i = 0^-] + \sum_{i=1}^{n} w_{i,+}(h)^2 \mathbb{V}[M_i(\bar{\mu})|X_i = 0^+] + o_p((nh)^{-1}).$$

It then follows from standard local linear arguments, that $nh\,\mathrm{se}^2(h;\bar{\mu}) - V(\bar{\mu}) = o_P(1)$ and $\mathbb{E}[\widehat{\tau}(h;\bar{\mu})|\mathbb{X}_n] - \tau = B(\bar{\mu})h^2 + o_p(h^2)$.

Part (iii). We prove this part by showing a more general result. Let

$$\begin{aligned}\widetilde{V}(\mu) &= \omega_+ \mathbb{V}[M_i(\mu)|X_i = 0^+] + \omega_- \mathbb{V}[M_i(\mu)|X_i = 0^-],\\ \mu_0^*(z) &= \frac{\omega_-}{\omega_- + \omega_+}\mu_0^-(z) + \frac{\omega_+}{\omega_- + \omega_+}\mu_0^+(z).\end{aligned}$$

We will show that for all $\mu^{(a)}, \mu^{(b)} \in \mathcal{M}$, it holds that $\widetilde{V}(\mu^{(a)}) < \widetilde{V}(\mu^{(b)})$ if and only if $\mathbb{V}[\mu^{(a)}(Z_i) - \mu_0^*(Z_i)|X_i = 0] < \mathbb{V}[\widetilde{\mu}^{(b)}(Z_i) - \mu_0^*(Z_i)|X_i = 0]$. The statement from Theorem 1 then follows by setting $\omega_+ = \omega_-$.

Fix $\mu \in \mathcal{M}$. By basic properties of the conditional expectation, we have that

$$\widetilde{V}(\mu) = \omega_+ \mathbb{V}[Y_i - \mu_0^+(Z_i)|X_i = 0^+] + \omega_- \mathbb{V}[Y_i - \mu_0^-(Z_i)|X_i = 0^-] + \widetilde{R}(\mu),$$

where the first two terms on the right-hand side do not depend on μ, and

$$\widetilde{R}(\mu) = \omega_+ \mathbb{V}[\mu_0^+(Z_i) - \mu(Z_i)|X_i = 0] + \omega_- \mathbb{V}[\mu_0^-(Z_i) - \mu(Z_i)|X_i = 0].$$

Further, it holds that

$$\begin{aligned}\widetilde{R}(\mu) = \widetilde{R}(\mu_0^* + \mu - \mu_0^*) &= \omega_+ \mathbb{V}\left[\frac{\omega_-}{\omega_+ + \omega_-}(\mu_0^+(Z_i) - \mu_0^-(Z_i)) - (\mu(Z_i) - \mu_0^*(Z_i))|X_i = 0\right]\\ &\quad + \omega_- \mathbb{V}\left[\frac{-\omega_+}{\omega_+ + \omega_-}(\mu_0^+(Z_i) - \mu_0^-(Z_i)) - (\mu(Z_i) - \mu_0^*(Z_i))|X_i = 0\right]\\ &= \widetilde{R}(\mu_0^*) + (\omega_+ + \omega_-)\mathbb{V}[\mu(Z_i) - \mu_0^*(Z_i)|X_i = 0],\end{aligned}$$

which concludes this proof. □

B. STANDARD ERROR

To estimate the variance of our estimator, we use a standard error of the form

$$\widehat{se}^2_{CF}(h;\widehat{\mu}) = \sum_{i=1}^{n} w_i^2(h)\widehat{\sigma}_i^2(\widehat{\mu}_{s(i)}),$$

where $\widehat{\sigma}_i^2(\widehat{\mu}_{s(i)})$ is an estimator of the variance $\sigma_i^2(\bar{\mu}) = \mathbb{V}[M_i(\bar{\mu})|X_i]$. Following Noack and Rothe (2021), we consider a version of the nearest neighbor variance estimator of Abadie et al. (2014).[8] We choose some R, say $R = 5$, which determines the number of neighbors to be used in the variance estimation. Based on the realized running variable, for each unit i, we determine its R nearest neighbors that are on the same side of the cutoff and within the same fold as unit i. Our estimator $\widehat{\sigma}_i^2(\widehat{\mu}_{s(i)})$ is proportional to the squared difference between $M_i(\widehat{\mu}_{s(i)})$ and its best linear predictor given the running variable based on its R nearest neighbors. We first introduce the notation. Let $\mu \in \mathcal{M}_n$. We denote the standard error by $\widehat{\text{se}}^2(h;\mu) = \sum_{i=1}^n w_i^2(h)\widehat{\sigma}_i^2(\mu)$, where

$$\widehat{\sigma}_i^2(\mu) = \frac{1}{1+H_i}\left(M_i(\mu) - \sum_{j\in\mathcal{R}_i} v_{j,i}M_j(\mu)\right)^2,$$

$$v_{j,i} = \widetilde{X}_i\left(\sum_{j\in\mathcal{R}_i}\widetilde{X}_j^\top\widetilde{X}_j\right)^{-1}\widetilde{X}_j^\top, \quad H_i = \widetilde{X}_i\left(\sum_{j\in\mathcal{R}_i}\widetilde{X}_j^\top\widetilde{X}_j\right)^{-1}\widetilde{X_i}$$

Here $\widetilde{X}_i = (1, X_i)$ and $\mathcal{R}_i$ is the set of the R nearest neighbors of unit i based on the running variable and within the same fold and on the same side of the cutoff as unit i. We note that by basic OLS algebra, the weights $v_{j,i}$ satisfy: $\sum_{j\in\mathcal{R}_i} v_{j,i} = 1$, $\sum_{j\in\mathcal{R}_i} v_{j,i}(X_j - X_i) = 0$, and $\sum_{j\in\mathcal{R}_i} v_{j,i}^2 = H_i$.

The following proposition establishes its consistency under standard regularity conditions.

Proposition B.1. *Suppose that Assumptions 1–5 hold and that for all $x \in \mathcal{X}$ and $n \in \mathbb{N}$, $\sup_{\mu\in\mathcal{T}_n}\mathbb{E}[(M_i(\mu) - \mathbb{E}[M_i(\mu)|X_i])^4|X_i = x]$ is bounded by B. Then*

$$nh\,\widehat{se}_{CF}^2(h;\widehat{\mu}) - V(\bar{\mu}) = o_P(1).$$

We further let $\widehat{\text{se}}_s^2(h;\mu) = \sum_{i\in I_s} w_i^2(h)\widehat{\sigma}_i^2(\mu)$, so that $\widehat{\text{se}}^2(h;\mu) = \sum_{s=1}^S \widehat{\text{se}}_s^2(h;\mu)$. Similarly, we define $\text{se}_s^2(h;\mu) = \sum_{i\in I_s} w_i^2(h)\sigma_i^2(\mu)$ and $\text{se}^2(h;\mu) = \sum_{s=1}^S \text{se}_s^2(h;\mu)$.

Proof of Proposition B.1. Using the triangular inequality, we first note that

$$\begin{aligned}|nh\,\widehat{\text{se}}_{CF}^2(h;\widehat{\mu}_n) - V(\bar{\mu})| &\leq nh|\widehat{\text{se}}_{CF}^2(h;\widehat{\mu}_n) - \text{se}^2(h;\bar{\mu})| + |nh\,\text{se}^2(h;\bar{\mu}) - V(\bar{\mu})|\\ &\leq S\max_{s\in[S]} nh|\widehat{\text{se}}_s^2(h;\widehat{\mu}_s) - \text{se}_s^2(h;\bar{\mu})| + o_p(1),\end{aligned}$$

[8] Alternatively, one can use the Eicker-Huber-White (EHW) standard error, but it might be conservative in finite samples; see the discussion by Abadie et al. (2014) in the standard nonparametric regression context.

where the second inequality follows from the proof of Theorem 1. The main step in this proof is to show that for any $s \in [S]$ and conditional on $\mathbb{X}_n$ and $(W_j)_{j\in I_s^c}$, it holds that

$$nh|\widehat{\text{se}}_s^2(h;\widehat{\mu}_s) - \text{se}_s^2(h;\bar{\mu})| = o_P(1). \tag{B.1}$$

We remark that the condition in (B.1) would essentially follow from the results of Noack and Rothe (2021) if $\mathbb{V}[M_i(\mu)|X_i = x]$ was L-Lipschitz continuous for all $\mu \in \mathcal{T}_n$. Our setting is different as we impose L-Lipschitz continuity only for the function $\mathbb{V}[M_i(\bar{\mu})|X_i = x]$. Still, some steps of our proof follow from the proof of Theorem 4 of Noack and Rothe (2021). We note that

$$\begin{aligned}
&\widehat{\text{se}}_s^2(h;\widehat{\mu}_s) - \text{se}_s^2(h;\bar{\mu}) \\
&= (\mathbb{E}[\widehat{\text{se}}_s^2(h;\bar{\mu})|\mathbb{X}_n] - \text{se}_s^2(h;\bar{\mu})) + (\widehat{\text{se}}_s^2(h;\widehat{\mu}_s) - \mathbb{E}[\widehat{\text{se}}_s^2(h;\widehat{\mu}_s)|\mathbb{X}_n,(W_j)_{j\in I_s^c}]) \\
&\quad + (\mathbb{E}[\widehat{\text{se}}_s^2(h;\widehat{\mu}_s) - \widehat{\text{se}}_s^2(h;\bar{\mu})|\mathbb{X}_n,(W_j)_{j\in I_s^c}]) \\
&\equiv G_1 + G_2 + G_3.
\end{aligned}$$

In the following, we show that each of the three terms is of order $o_P((nh)^{-1})$. First, it follows from the proof of Theorem 4 of Noack and Rothe (2021) that $G_1 = o_P((nh)^{-1})$ as $\mathbb{V}[M_i(\bar{\mu})|X_i = x]$ is L-Lipschitz continuous by Assumption 5.

Second, it is clear that $\mathbb{E}[G_2|\mathbb{X}_n,(W_j)_{j\in I_s^c}] = 0$. Further, it follows that with probability approaching one,

$$\mathbb{E}[G_2^2|\mathbb{X}_n,(W_j)_{j\in I_s^c}] \leq \sup_{\mu\in\mathcal{T}_n} \mathbb{E}\left[\left(\widehat{\text{se}}_s^2(h;\mu) - \mathbb{E}[\widehat{\text{se}}_s^2(h;\mu)|\mathbb{X}_n]\right)^2\right] = o_p((nh)^{-2}),$$

where the last equality follows from the proof of Theorem 4 of Noack and Rothe (2021) using boundedness of the fourth conditional moment assumed in the proposition.

We now consider G_3. We note that with probability approaching one

$$\begin{aligned}
|G_3| &= |\sum_{i\in I_s} w_i^2(h)\mathbb{E}[\widehat{\sigma}_i^2(\widehat{\mu}_s) - \widehat{\sigma}_i^2(\bar{\mu})|\mathbb{X}_n,(W_j)_{j\in I_s^c}]| \\
&\leq \sup_{j\in I_s\colon X_j\in\mathcal{X}_h} \sup_{\mu\in\mathcal{T}_n} \left|\mathbb{E}[\widehat{\sigma}_j^2(\mu) - \widehat{\sigma}_j^2(\bar{\mu})|\mathbb{X}_n]\right| \sum_{i\in I_s} w_i(h)^2.
\end{aligned}$$

Following Noack and Rothe (2021), we note that for any $\mu \in \mathcal{T}_n$ and any $i \in I_s$

$$\begin{aligned}\mathbb{E}[\widehat{\sigma}_i(\mu)|\mathbb{X}_n] = \sigma_i^2(\mu) + \frac{1}{1+H_i}\left(\sum_{j \in \mathcal{R}_i} v_{j,i}^2(\sigma_j^2(\mu) - \sigma_i^2(\mu))\right) \qquad \text{(B.2)}\\ + \frac{1}{1+H_i}\left(\mathbb{E}[M_i(\mu)|X_i] - \sum_{j \in \mathcal{R}_i} v_{j,i}\mathbb{E}[M_j(\mu)|X_j]\right)^2.\end{aligned}$$

In the following, we denote by C a positive constant, which might be different from line to line. By a second-order Taylor-expansion and by a simple OLS-algebra, it holds for the last term in the above expression that

$$\begin{aligned}\sup_{i \in I_s\colon X_i \in \mathcal{X}_h} \sup_{\mu \in \mathcal{T}_n} \frac{1}{1+H_i}\left(\mathbb{E}[M_i(\mu)|X_i] - \sum_{j \in \mathcal{R}_i} v_{j,i}\mathbb{E}[M_j(\mu)|X_j]\right)^2 \qquad \text{(B.3)}\\ \leq C \sup_{i \in I_s\colon X_i \in \mathcal{X}_h} \sup_{j \in \mathcal{R}_i} |X_i - X_j|^4 \sup_{x \in \mathcal{X}_h} \sup_{\mu \in \mathcal{T}_n} (\partial_x^2 \mathbb{E}[M_i(\mu)|X_i = x])^2 = o_p(1),\end{aligned}$$

where we used that $\frac{1}{1+H_i}\sum_{j \in \mathcal{R}_i} v_{j,i}^2 \leq 1$ and $\sup_{x \in \mathcal{X}_h} \sup_{\mu \in \mathcal{T}_n} \partial_x^2 \mathbb{E}[M_i(\mu)|X_i = x] = O(1)$ by Assumptions 4 and 5.

Using (B.2) and (B.3), we obtain that

$$\begin{aligned}&\sup_{i \in I_s\colon X_i \in \mathcal{X}_h} \sup_{\mu \in \mathcal{T}_n} |\mathbb{E}[\widehat{\sigma}_i^2(\mu) - \widehat{\sigma}_i^2(\bar{\mu})|\mathbb{X}_n]|\\ &\leq \sup_{i \in I_s\colon X_i \in \mathcal{X}_h} \sup_{\mu \in \mathcal{T}_n} \left|\sigma_i^2(\mu) - \sigma_i^2(\bar{\mu}) + \frac{1}{1+H_i}\left(\sum_{j \in \mathcal{R}_i} v_{j,i}^2(\sigma_j^2(\mu) - \sigma_j^2(\bar{\mu}) + \sigma_i^2(\bar{\mu}) - \sigma_i^2(\mu))\right)\right| + o_p(1)\\ &\leq C \sup_{i \in I_s\colon X_i \in \mathcal{X}_h} \sup_{\mu \in \mathcal{T}_n} |\sigma_i^2(\mu) - \sigma_i^2(\bar{\mu})| + o_p(1)\\ &\leq C \sup_{x \in \mathcal{X}_h} \sup_{\mu \in \mathcal{T}_n} |\mathbb{V}[M_i(\mu)|X_i = x] - \mathbb{V}[M_i(\bar{\mu})|X_i = x]| + o_p(1)\\ &= o_p(1),\end{aligned}$$

where we used that $\frac{1}{1+H_i}\sum_{j \in \mathcal{R}_i} v_{j,i}^2 \leq 1$ and Assumption 3. Since $\sum_{i \in I_s} w_i(h)^2 = O_p((nh)^{-1})$, we conclude that $G_3 = o_P((nh)^{-1})$.

C. ADDITIONAL SIMULATION RESULTS

In this section, we present further simulation results. Table C.1 extends the results in Table 1. Apart from the bias-aware approach discussed in the main text, we consider bandwidth choices and confidence intervals based on robust bias corrections and undersmoothing (the bandwidth for undersmoothing is chosen as $n^{-1/20}$ times the MSE-optimal bandwidth estimated using the `rdrobust` package). The qualitative conclusions about the relative performance of different first-stage estimators in different models remain the same as discussed in the main text.

The simulated average bandwidth of robust bias corrections is on typically smaller than that of the bias-aware approach, and the confidence intervals are larger. This feature is known in the nonparametric literature. In the last two rows of Table C.1, we report the results using the procedure of Calonico et al. (2019). In this simulation setting, they are essentially the same as the results for our procedure with a linear adjustment function.

In Table C.2, we report simulation results for DGP 3 for different values of the signal-to-noise ratio. This illustrates that the potential gains from covariate adjustments are large if the additional covariates explain a large portion of variation in the outcome variable.

Table C.1: Full simulation results for different numbers of relevant covariates

		Cov	Bias	SD	CI	h	Cov	Bias	SD	CI	h	Cov	Bias	SD	CI	h	Cov	Bias	SD	CI	h
		Model 1: L=0					**Model 2: L=4**					**Model 3: L=10**					**Model 4: L=25**				
Standard	BA	0.970	-0.014	0.074	0.324	0.432	0.961	-0.071	0.186	0.818	0.688	0.964	-0.095	0.191	0.876	0.793	0.959	-0.063	0.185	0.810	0.685
	RBC	0.948	0.015	0.110	0.415	0.299	0.947	0.000	0.351	1.309	0.305	0.946	0.011	0.371	1.400	0.269	0.942	0.016	0.393	1.457	0.245
	US	0.949	0.006	0.113	0.425	0.205	0.945	-0.011	0.360	1.334	0.209	0.947	0.001	0.380	1.424	0.184	0.943	0.008	0.405	1.484	0.167
Optimal Inf	BA	0.970	-0.014	0.074	0.324	0.432	0.966	-0.015	0.075	0.325	0.432	0.965	-0.013	0.076	0.325	0.432	0.969	-0.013	0.074	0.324	0.432
	RBC	0.948	0.015	0.110	0.415	0.299	0.943	0.013	0.110	0.415	0.299	0.936	0.015	0.113	0.415	0.299	0.942	0.015	0.110	0.414	0.300
	US	0.949	0.006	0.113	0.425	0.205	0.945	0.003	0.113	0.425	0.204	0.937	0.005	0.116	0.426	0.204	0.944	0.005	0.113	0.425	0.205
Linear Inf	BA	0.970	-0.014	0.074	0.324	0.432	0.966	-0.015	0.075	0.325	0.432	0.967	-0.048	0.127	0.562	0.618	0.968	-0.043	0.103	0.472	0.590
	RBC	0.948	0.015	0.011	0.415	0.299	0.943	0.013	0.011	0.415	0.299	0.937	0.013	0.234	0.859	0.263	0.946	0.007	0.199	0.754	0.197
	US	0.949	0.006	0.113	0.425	0.205	0.945	0.003	0.113	0.425	0.204	0.941	0.003	0.239	0.876	0.180	0.942	0.002	0.205	0.766	0.134
Linear	BA	0.970	-0.014	0.074	0.327	0.433	0.967	-0.015	0.075	0.326	0.433	0.959	-0.040	0.137	0.591	0.597	0.965	-0.043	0.108	0.492	0.588
	RBC	0.948	0.015	0.110	0.418	0.300	0.943	0.014	0.111	0.418	0.299	0.940	0.016	0.250	0.918	0.279	0.943	0.007	0.216	0.810	0.203
	US	0.951	0.006	0.113	0.429	0.205	0.946	0.003	0.114	0.428	0.205	0.942	0.006	0.256	0.937	0.191	0.944	0.002	0.222	0.824	0.139
Local Linear	BA	0.970	-0.014	0.074	0.327	0.433	0.968	-0.014	0.075	0.327	0.433	0.963	-0.016	0.083	0.356	0.452	0.968	-0.016	0.082	0.359	0.456
	RBC	0.945	0.015	0.111	0.419	0.300	0.945	0.014	0.111	0.419	0.299	0.943	0.014	0.127	0.471	0.293	0.942	0.015	0.130	0.490	0.278
	US	0.949	0.006	0.114	0.429	0.205	0.947	0.004	0.114	0.430	0.205	0.943	0.005	0.131	0.482	0.200	0.943	0.005	0.135	0.501	0.190
Lasso	BA	0.967	-0.014	0.076	0.331	0.436	0.966	-0.021	0.088	0.383	0.466	0.962	-0.020	0.092	0.391	0.467	0.968	-0.014	0.077	0.340	0.443
	RBC	0.944	0.015	0.116	0.435	0.288	0.950	0.012	0.138	0.521	0.291	0.939	0.013	0.146	0.530	0.295	0.943	0.011	0.132	0.490	0.243
	US	0.951	0.007	0.118	0.445	0.197	0.947	0.002	0.142	0.532	0.199	0.941	0.004	0.150	0.542	0.202	0.942	0.005	0.135	0.500	0.166
Forest	BA	0.968	-0.015	0.076	0.331	0.436	0.967	-0.021	0.087	0.379	0.465	0.966	-0.019	0.085	0.372	0.469	0.971	-0.022	0.093	0.413	0.490
	RBC	0.946	0.015	0.113	0.425	0.299	0.949	0.010	0.134	0.507	0.297	0.940	0.010	0.152	0.560	0.233	0.940	0.008	0.186	0.688	0.198
	US	0.946	0.006	0.116	0.436	0.205	0.948	0.000	0.138	0.518	0.203	0.943	0.003	0.155	0.570	0.159	0.943	0.004	0.191	0.701	0.136
CCFT	RBC	0.945	0.014	0.110	0.413	0.297	0.939	0.013	0.111	0.413	0.297	0.934	0.013	0.235	0.851	0.263	0.943	0.007	0.201	0.746	0.196
	US	0.944	0.006	0.114	0.422	0.203	0.940	0.003	0.114	0.422	0.203	0.934	0.003	0.241	0.863	0.180	0.935	0.002	0.208	0.752	0.134

Notes: Results based on 5000 Monte Carlo draws based on Model 3 explained in the main text. Columns show results for simulated coverage for a nominal confidence level of 95% (Cov); the mean bias (Bias); the mean Standard Deviation (SD); the mean confidence interval length (CI); and the mean bandwidth (h). Bandwidth and confidence intervals are constructed based on the bias-aware approach (BA), robust bias correction (RBC), and undersmoothing (US).

Table C.2: Simulation results for different signal-to-noise ratios

		Cov	Bias	SD	CI	h	Cov	Bias	SD	CI	h	Cov	Bias	SD	CI	h	Cov	Bias	SD	CI	h
		$\rho = 3$					$\rho = 1$					$\rho = 5$					$\rho = 10$				
Standard	BA	96.2	-8.9	19.6	87.5	79.3	96.4	-3.1	9.8	43.3	52.0	96.0	-14.7	29.6	134.7	95.5	95.5	-16.8	58.6	241.5	99.9
	RBC	94.7	1.1	37.2	139.6	26.9	94.1	0.8	16.4	61.2	27.9	94.6	0.9	59.8	226.5	26.7	94.7	-0.1	118.3	446.9	26.7
	US	94.3	0.4	38.2	142.1	18.4	93.8	-0.2	16.9	62.4	19.1	94.5	-0.2	61.4	230.3	18.3	94.8	-0.5	120.5	454.4	18.3
Optimal Inf	BA	97.0	-1.4	7.4	32.4	43.2	96.6	-1.5	7.4	32.5	43.2	96.5	-1.3	7.6	32.4	43.2	96.9	-1.3	7.3	32.5	43.2
	RBC	94.8	1.5	11.0	41.5	29.9	94.3	1.3	11.0	41.5	29.9	93.5	1.5	0.113	41.5	29.8	94.2	1.5	11.0	41.4	30.0
	US	94.9	0.6	0.113	42.5	20.5	94.5	0.3	0.113	42.5	20.4	93.6	0.6	11.6	42.6	20.4	94.5	0.5	0.113	42.5	20.5
Linear Inf	BA	96.0	-4.8	12.9	56.2	61.9	96.2	-1.9	8.3	36.1	46.1	96.4	-8.5	18.1	81.5	76.6	0.959	-14.1	33.0	146.9	96.3
	RBC	94.2	1.2	23.1	85.8	26.4	94.0	1.2	13.1	48.9	28.3	93.8	1.2	35.8	131.6	25.8	94.8	1.2	66.2	252.9	25.6
	US	93.9	0.6	23.8	87.5	18.0	94.4	0.2	13.4	49.9	19.4	94.1	0.1	36.6	134.0	17.7	95.0	0.6	67.2	257.4	17.5
Linear	BA	96.1	-4.0	13.8	59.1	59.8	96.1	-1.9	8.4	36.5	45.9	95.6	-7.2	21.2	90.7	74.0	0.958	-14.1	37.0	161.5	95.4
	RBC	94.0	1.7	24.9	91.9	27.9	94.0	1.2	13.2	49.3	28.6	94.0	1.7	42.2	155.3	28.8	94.6	2.3	83.4	312.8	29.0
	US	93.9	1.0	25.6	93.8	19.1	94.3	0.3	13.5	50.4	19.5	94.2	0.7	43.3	158.4	19.7	94.8	1.0	85.5	318.8	19.8
Local Linear	BA	96.7	-1.7	8.2	35.6	45.1	96.5	-1.5	7.8	33.9	44.1	96.5	-1.7	8.7	37.3	46.3	97.0	-2.6	10.2	45.1	52.3
	RBC	94.4	1.5	12.5	47.1	29.3	94.2	1.3	11.7	43.9	29.7	94.0	1.5	13.7	50.5	28.8	94.6	1.6	17.4	65.1	27.6
	US	94.7	0.7	12.9	48.2	20.1	94.3	0.4	12.0	45.0	20.3	94.0	0.6	14.1	51.7	19.7	94.6	0.7	18.0	66.4	18.9
Lasso	BA	96.8	-2.0	9.1	39.3	46.9	96.8	-1.6	7.7	34.0	44.5	96.1	-2.7	11.5	48.4	51.0	96.2	-4.9	18.1	75.8	61.6
	RBC	93.8	1.5	14.4	53.3	29.6	94.3	1.0	12.4	46.9	26.5	93.9	1.5	18.5	67.7	31.1	94.1	1.9	32.6	117.7	32.2
	US	94.3	0.6	14.9	54.6	20.2	94.4	0.3	12.7	47.9	18.1	94.2	0.6	19.1	69.2	21.3	94.2	0.8	33.3	120.1	22.0
Forest	BA	96.6	-1.9	8.5	37.2	46.9	96.5	-1.6	7.7	33.7	44.3	96.7	-2.6	10.0	43.8	51.1	96.3	-5.8	14.7	64.5	64.5
	RBC	94.1	1.1	15.1	56.1	23.1	94.1	1.1	12.1	45.3	27.7	94.5	0.8	18.9	70.6	21.8	93.9	0.5	31.7	116.1	20.7
	US	94.3	0.6	15.5	57.2	15.8	94.5	0.2	12.4	46.2	19.0	95.0	0.2	19.3	71.8	14.9	94.2	0.1	32.5	118.1	14.2
CCFT	RBC	93.8	1.2	23.3	85.0	26.3	93.5	1.1	13.2	48.6	28.1	93.4	1.2	36.0	130.3	25.8	94.6	1.3	66.2	250.4	25.6
	US	93.3	0.6	24.0	86.3	18.0	93.9	0.2	13.5	49.4	19.2	93.3	0.2	36.9	132.0	17.7	94.7	0.5	67.4	253.4	17.5

Notes: Results based on 5000 Monte Carlo draws based on Model 3 explained in the main text. All numbers are multiplied by 100. Columns show results for simulated coverage for a nominal confidence level of 95% (Cov); the mean bias (Bias); the mean Standard Deviation (SD); the mean confidence interval length (CI); and the mean bandwidth (h). Bandwidth and confidence intervals are constructed based on the bias-aware approach (BA), robust bias correction (RBC), and undersmoothing (US).

REFERENCES

Abadie, A., G. W. Imbens, and F. Zheng (2014): "Inference for misspecified models with fixed regressors," *Journal of the American Statistical Association*, 109, 1601–1614.

Andrews, D. (1994): "Asymptotics for semiparametric econometric models via stochastic equicontinuity," *Econometrica*, 62, 43–72.

Armstrong, T. B. and M. Kolesár (2018): "Optimal inference in a class of regression models," *Econometrica*, 86, 655–683.

——— (2020): "Simple and honest confidence intervals in nonparametric regression," *Quantitative Economics*, 11, 1–39.

Belloni, A., V. Chernozhukov, I. Fernández-Val, and C. Hansen (2017): "Program Evaluation and Causal Inference With High-Dimensional Data," *Econometrica*, 85, 233–298.

Calonico, S., M. D. Cattaneo, M. H. Farrell, and R. Titiunik (2019): "Regression Discontinuity Designs Using Covariates," *The Review of Economics and Statistics*, 101, 442–451.

Calonico, S., M. D. Cattaneo, and R. Titiunik (2014): "Robust nonparametric confidence intervals for regression-discontinuity designs," *Econometrica*, 82, 2295–2326.

Cattaneo, M. D., N. Idrobo, and R. Titiunik (2019): *A practical introduction to regression discontinuity designs: Foundations*, Cambridge University Press.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018): "Double/debiased machine learning for treatment and structural parameters," *The Econometrics Journal*, 21, C1–C68.

Chernozhukov, V., W. Newey, J. Robins, and R. Singh (2019): "Double/de-biased machine learning of global and local parameters using regularized Riesz representers," *Working Paper.*

Dong, Y. (2017): "Alternative Assumptions to Identify LATE in Fuzzy Regression Discontinuity Designs," *Working Paper.*

Fan, J. and I. Gijbels (1996): *Local polynomial modelling and its applications*, Chapman & Hall/CRC.

Fan, Q., Y.-C. Hsu, R. P. Lieli, and Y. Zhang (2020): "Estimation of Conditional Average Treatment Effects With High-Dimensional Data," *Journal of Business & Economic Statistics*, 0, 1–15.

Frölich, M. and M. Huber (2019): "Including Covariates in the Regression Discontinuity Design," *Journal of Business & Economic Statistics*, 37, 736–748.

Gelman, A. and G. Imbens (2019): "Why high-order polynomials should not be used in regression discontinuity designs," *Journal of Business & Economic Statistics*, 37, 447–456.

Gerard, F., M. Rokkanen, and C. Rothe (2020): "Bounds on treatment effects in regression discontinuity designs with a manipulated running variable," *Quantitative Economics*, 11, 839–870.

Hahn, J. (1998): "On the role of the propensity score in efficient semiparametric estimation of average treatment effects," *Econometrica*, 66, 315–331.

Hahn, J., P. Todd, and W. Van der Klaauw (2001): "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design," *Econometrica*, 69, 201–209.

Imbens, G. and K. Kalyanaraman (2012): "Optimal bandwidth choice for the regression discontinuity estimator," *Review of Economic Studies*, 79, 933–959.

Imbens, G. W. and T. Lemieux (2008): "Regression discontinuity designs: A guide to practice," *Journal of Econometrics*, 142, 615–635.

Kennedy, E. H. (2020): "Optimal doubly robust estimation of heterogeneous causal effects," *arXiv preprint arXiv:2004.14497*.

Kennedy, E. H., Z. Ma, M. D. McHugh, and D. S. Small (2017): "Nonparametric methods for doubly robust estimation of continuous treatment effects," *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 79, 1229.

Kreiss, A. and C. Rothe (2021): "Regression Discontinuity Analysis with Many Covariates," *Working Paper*.

Lee, D. S. and T. Lemieux (2010): "Regression discontinuity designs in economics," *Journal of Economic Literature*, 48, 281–355.

McCrary, J. (2008): "Manipulation of the running variable in the regression discontinuity design: A density test," *Journal of econometrics*, 142, 698–714.

Newey, W. (1994): "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, 62, 1349–1382.

Noack, C. and C. Rothe (2021): "Bias-aware inference in fuzzy regression discontinuity designs," *arXiv preprint arXiv:1906.04631*.

Robins, J. M. and A. Rotnitzky (2001): "Comment on "Inference for semiparametric models: some questions and an answer" by P. Bickel and J. Kwon," *Statistica Sinica*, 11, 920–936.

Wager, S., W. Du, J. Taylor, and R. J. Tibshirani (2016): "High-dimensional regression adjustments in randomized experiments," *Proceedings of the National Academy of Sciences*, 113, 12673–12678.